Detekcia komunít v komplexných sieťach

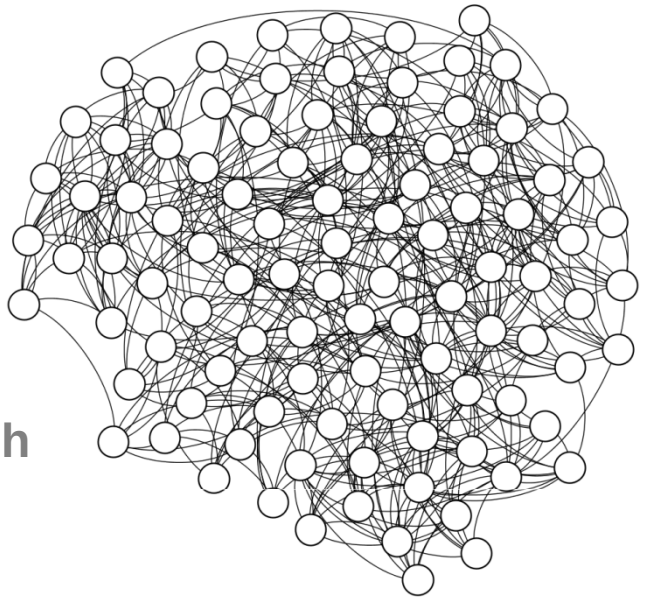
Marek Ciglan a Michal Laclavík,
Ústav informatiky SAV,
marek.ciglan@savba.sk, michal.laclavik@savba.sk

Kompetenčné centrum inteligentných technológií pre elektronizáciu a informatizáciu systémov a služieb
KC-INTELINSYS, ITMS: 26240220072

Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ.

Na realizáciu projektu sme získali nenávratný finančný príspevok v rámci Operačného programu výskum a vývoj.

Náčrt obsahu



- Úvod
  - **Komplexné siete v reálnom svete**
  - **Vlastnosti komplexných sietí**
    - Ako vyzerajú?
    - Porovnanie s náhodným sieťami.
  - **Typické úlohy dolovania v grafových dátach**
- Detekcia komunít
  - **Definícia úlohy**
    - Neexistencia konsenzu na formalizácii úlohy a dôsledky
    - Najrozšírenejšie špecifikácie problému používané v literatúre
  - **Triedy algoritmov na detekciu komunít**
- Meranie kvality zhlukovania
  - **Lokálne metriky kvality komunít**
  - **Modularita**

Náčrt obsahu

- Meranie kvality zhlukovania
  - **Moudlarita**
    - Popis
    - Problémy spojené s modularitou
  - **Benchmarky pre detekciu komunít**
    - GN benchmark
    - LFR benchmark
  - **Porovnanie kanonického a detekovaného rozdelenia**
    - NMI
    - Priemerná Jaccardová podobnosť
- Modifikácie základnej úlohy
  - **Perekrývajúce sa zkluky**
  - **Vrcholy / hrany s atribútmi**
  - **Detekcia komunity pre daný vrchol**

Náčrt obsahu

- Rýchle algoritmy na detekciu komunít
  - **Label propagation**
  - **Louvain method**
  - **SCCD**
- Zhlukovanie na sieťach reálneho sveta
  - **Ako dobre fungujú algoritmy na detekciu komunít na sieťach reálneho sveta s explicitne definovanými komunitami**

Komplexné siete

- **Grafové dáta** – zachytávajú entity (vrcholy grafu) a ich vzťahy (hrany grafu)
- **Veľké dáta** – rozsah spracovávaných dát robí orientáciu v dátach, ich vizualizáciu a štúdium ich vlastností netriviálnym problémom
- **Komplexný systém** – zložený zo spojených častí, ako celok vykazuje vlastnosti, ktoré sa nedajú pozorovať na úrovni jednotlivých častí

Komplexné siete v reálnom svete

- Čoraz viac dostupných dát, ktoré možno prirodzene modelovať ako komplexné siete
  - **Veľké on-line aplikácie:**
    - Sociálne siete (facebook, twitter)
    - Komunikácia (instant messenging, emaily, call networks)
    - Sociálne média (blogy)

Sociálne siete a obsah generovaný užívateľmi

- Sociálna sieť – relácie medzi užívateľmi
  - **Facebook - 750 miliónov užívateľov**
  - **YouTube - 490 miliónov pravidelných užívateľov**
  - **Twitter - 550 miliónov užívateľov**
  - **Wikipedia - 91,000 kontribútorov**
- Obsah generovaný užívateľmi
  - **textové správy, fotky, videá, reakcie (+1 / Likes)**
  - **Facebook - 30 miliárd zdieľaných položiek / mesiac**
  - **Twitter - 190 miliónov mikropostov / deň**
  - **Wikipedia - 17 miliónov hostovaných článkov**
- Interakcia užívateľov s obsahom
  - **YouTube - 92 miliárd zobrazení stránok / mesiac**
  - **Twitter - 1.6 miliárd dopytov za deň**
  - **Facebook – priemerný čas strávený na stránke za mesiac: 15H 33M**

Zdroj: http://www.jeffbullas.com/2011/09/02/20-stunning-social-media-tatistics/#q3eTJhr64rtD0tLF.99

Komplexné siete v reálnom svete

- Sociálne site
  - On-line sociálne siete
  - Komunikačné siete
- Informačné siete
  - Blogy
  - Citačné siete
  - WWW, hypertext
- Sémantické site
  - Linked open data cloud
- Jazykové siete
  - Term co-occurance networks
- Techologické siete
  - Cestné siete
  - Inžinierske site (elektrické, potrubné)

Vlastnosti komplexných sietí

- Ako vyzerajú komplexné siete z reálneho sveta?
- Ako sa podobajú reálne site na náhodné grafy?
- Majú rovnaké typy sietí podobné vlastnosti?
  - **Sú globálne vlastnosti sociálnej siete A podobné vlastnostiam sociálnej siete B?**

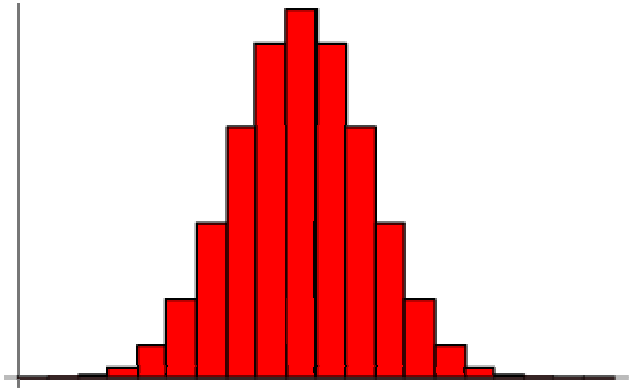
Štatistické vlastnosti komplexných sietí

- Podobajú sa komplexné siete na náhodné grafy?
- Erdos-Renyi model náhodného grafu:
  - **Variant 1:**
    - Graf o *n* vrcholoch, každá hrana *(i,j)* existuje s pravdepodobnosťou *p*
    - Teda graf s *m* hranami sa vyskytuje s pravdepodobnoťou:
    - *p^m * (1-p)^(M-m)*; kde *M=n(n-1)/2*
  - **Variant 2:**
    - Graf on n vrcholoch a m náhodne vybraných hranách

Štatistické vlastnosti komplexných sietí

- Podobajú sa komplexné siete na náhodné grafy?
- Erdos-Renyi model náhodného grafu:
  - **Variant 1:**
    - Graf o *n* vrcholoch, každá hrana *(i,j)* existuje s pravdepodobnosťou *p*
    - Teda graf s m hranami sa vyskytuje s pravdepodobnoťou:
    - *p^m * (1-p)^(M-m)*; kde *M=n(n-1)/2*
  - **Variant 2:**
    - Graf on n vrcholoch a m náhodne vybraných hranách

- Distribúcia stupňov náhodného grafu - binomická

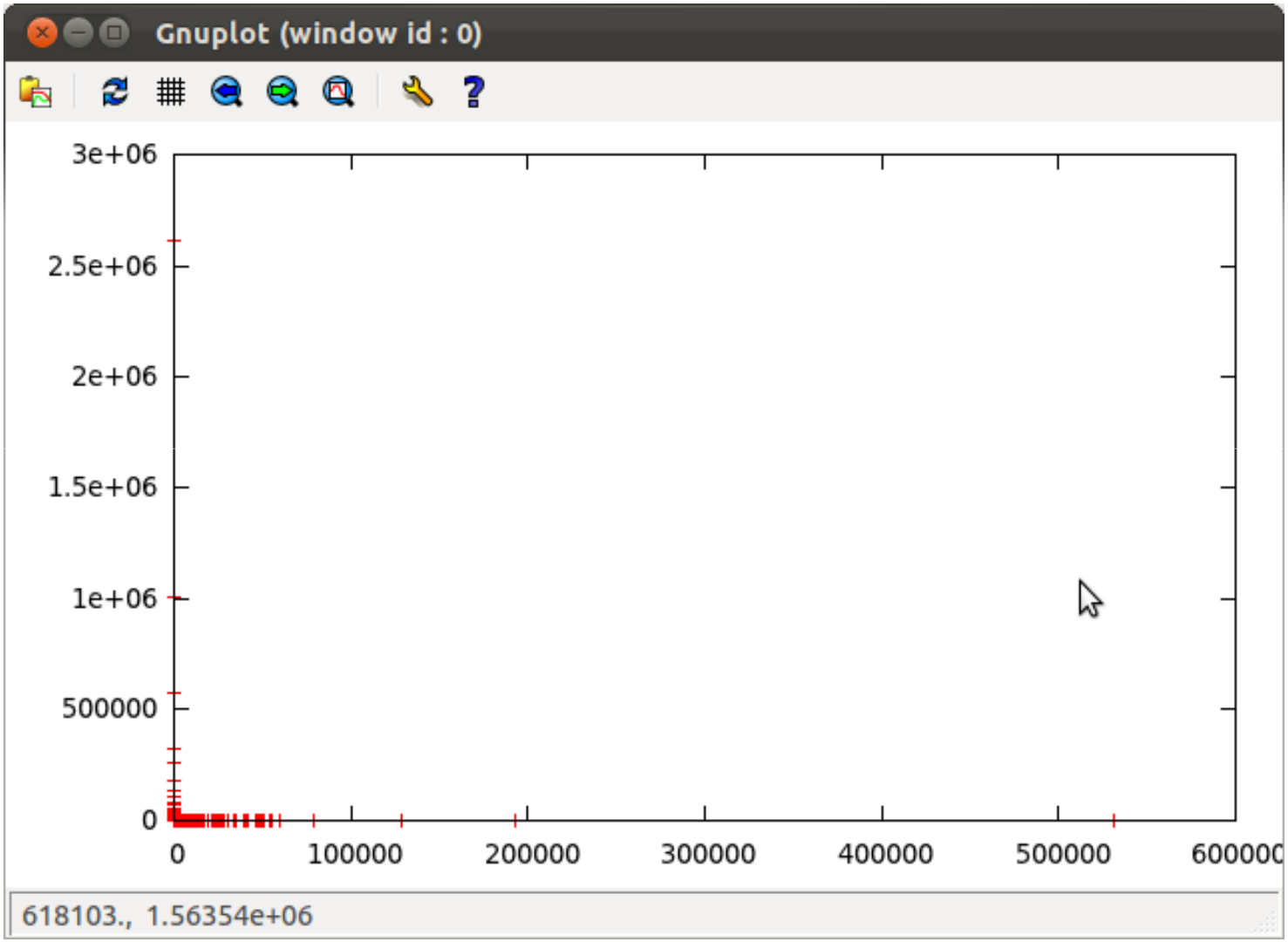


$$p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

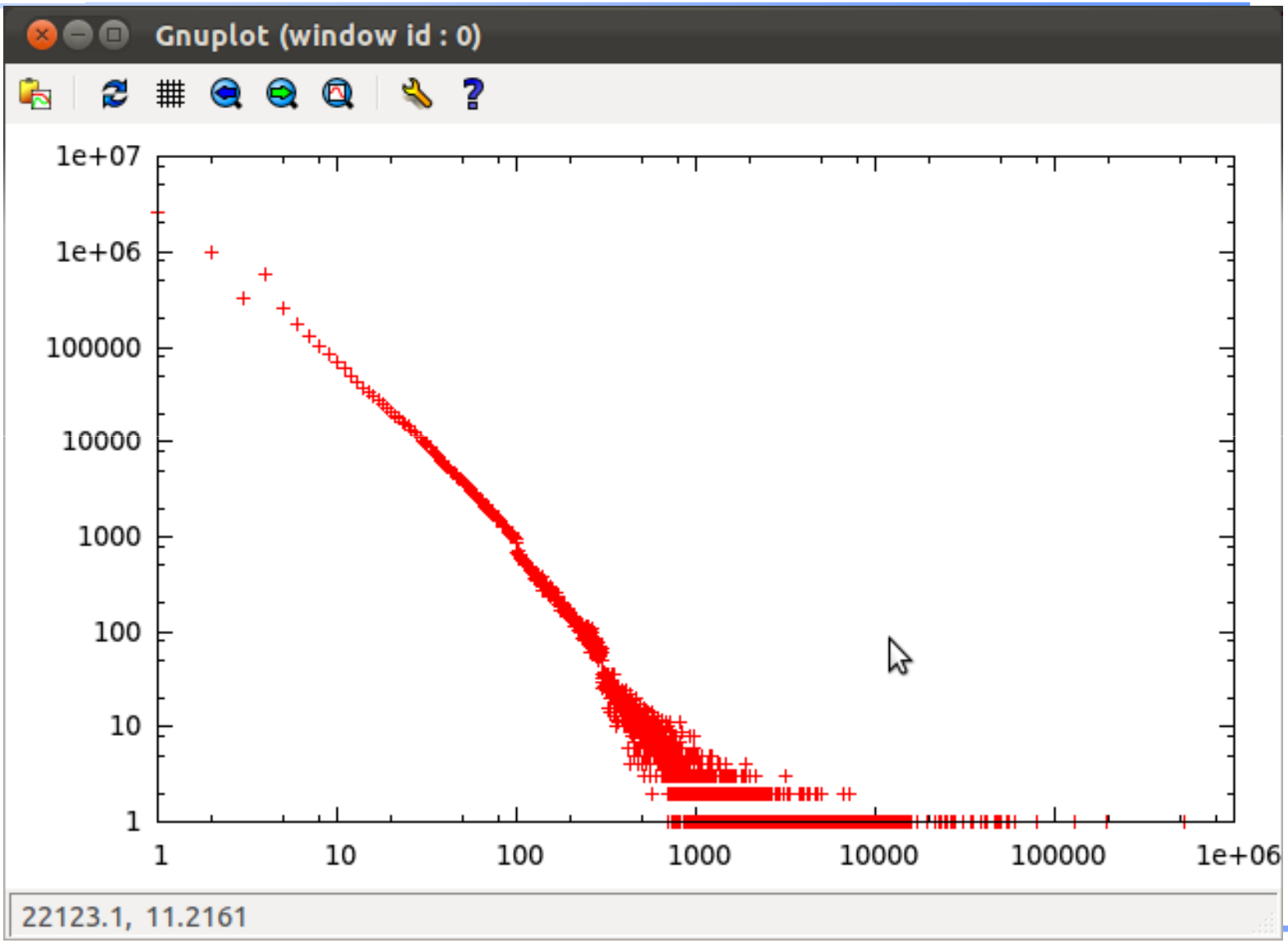
Štatistické vlastnosti komplexných sietí

- Podobajú sa komplexné siete na náhodné grafy?
- Erdos-Renyi model náhodného grafu:
  - **Variant 1:**
    - Graf o *n* vrcholoch, každá hrana *(i,j)* existuje s pravdepodobnosťou *p*
    - Teda graf s m hranami sa vyskytuje s pravdepodobnoťou:
    - *p^m * (1-p)^(M-m)*; kde *M=n(n-1)/2*
  - **Variant 2:**
    - Graf on n vrcholoch a m náhodne vybraných hranách
- Distribúcia stupňov náhodného grafu – binomická
- Priemer grafu: O(log n) - zväčšujúci sa s veľkosťou grafu

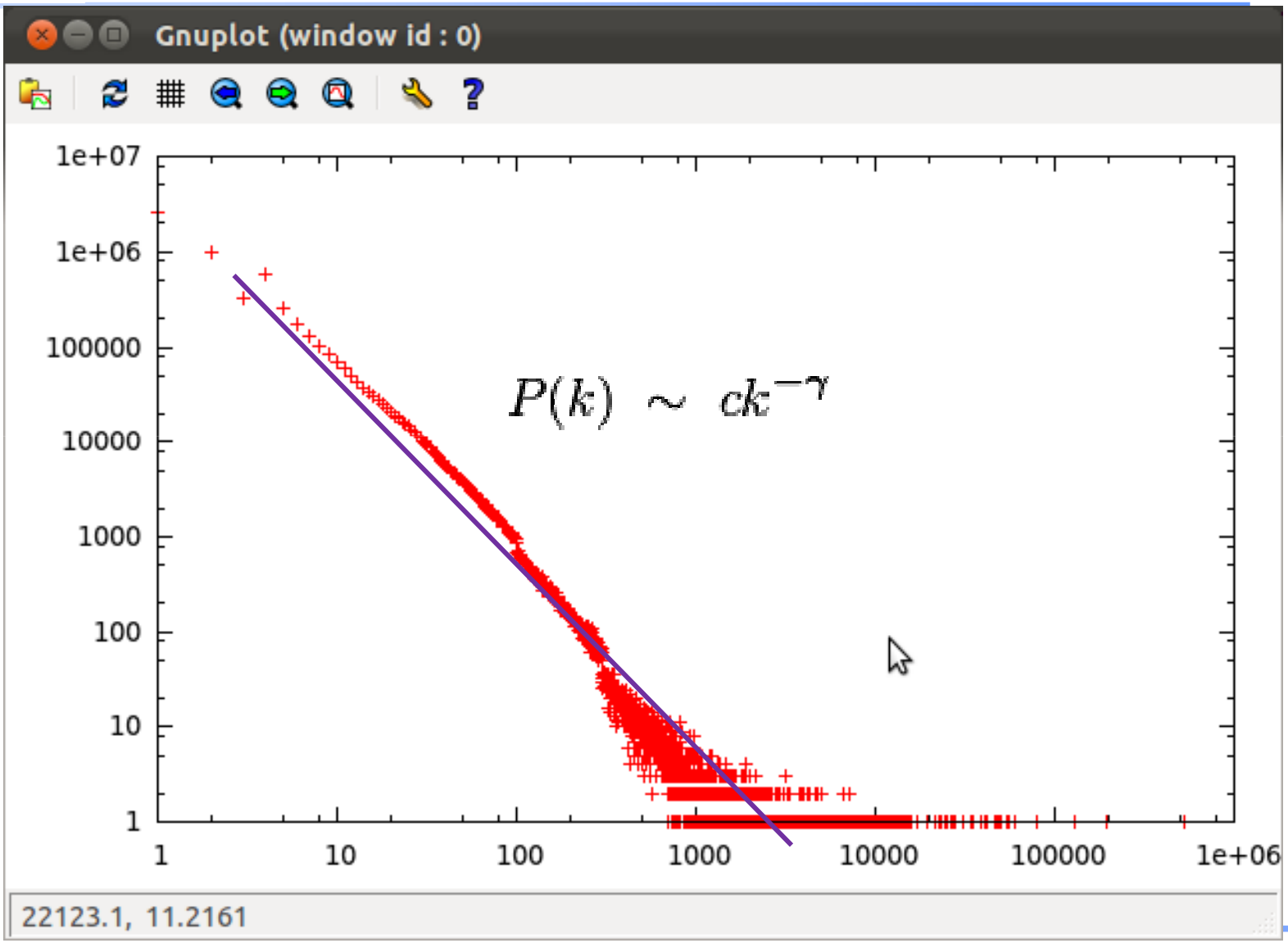
Distribúcia stupňov v reálnych sieťach (príklad DBpedia)



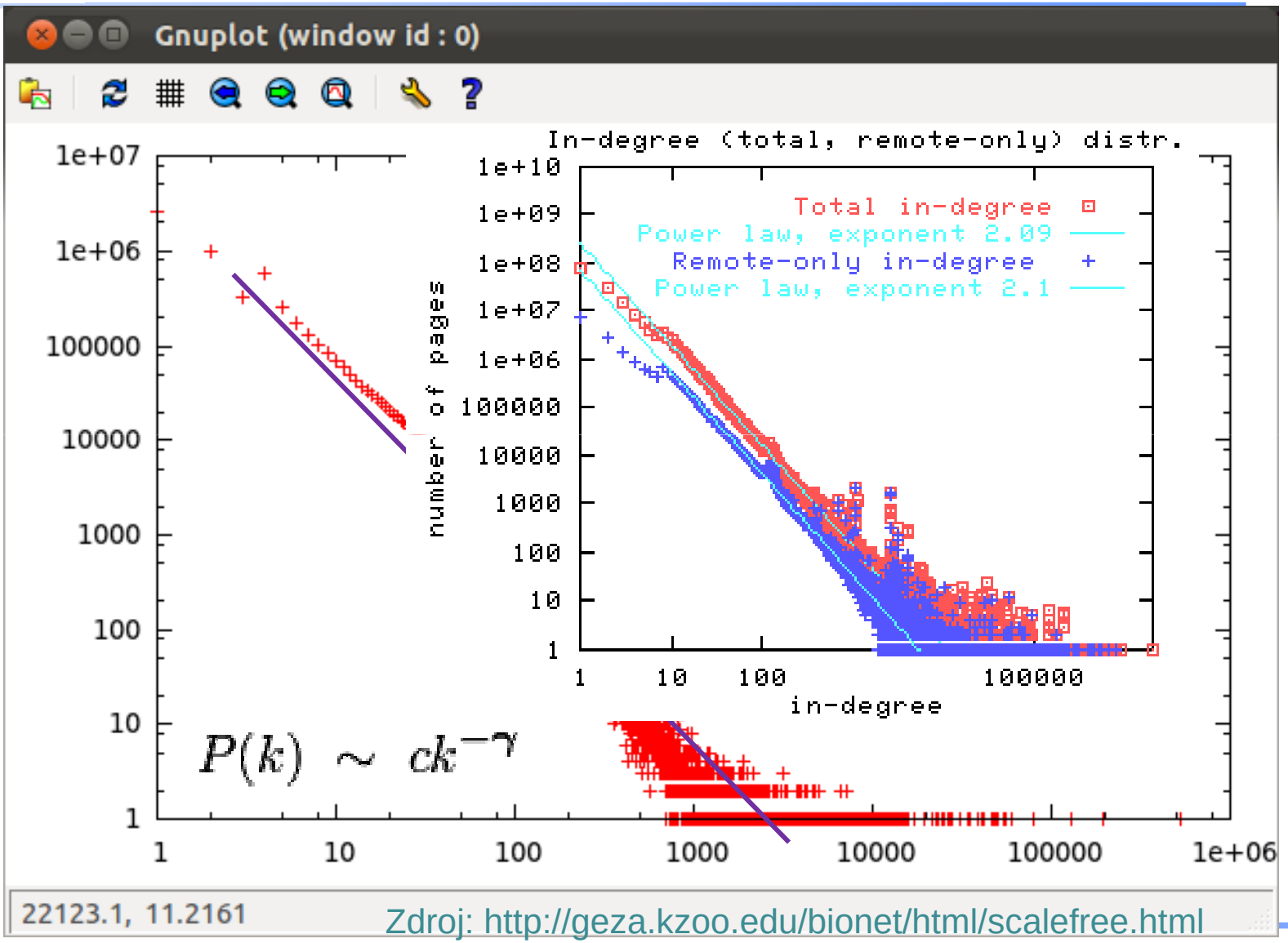
Distribúcia stupňov v reálnych sieťach - log škála (príklad DBpedia)



Distribúcia stupňov v reálnych sieťach - log škála (príklad DBpedia)



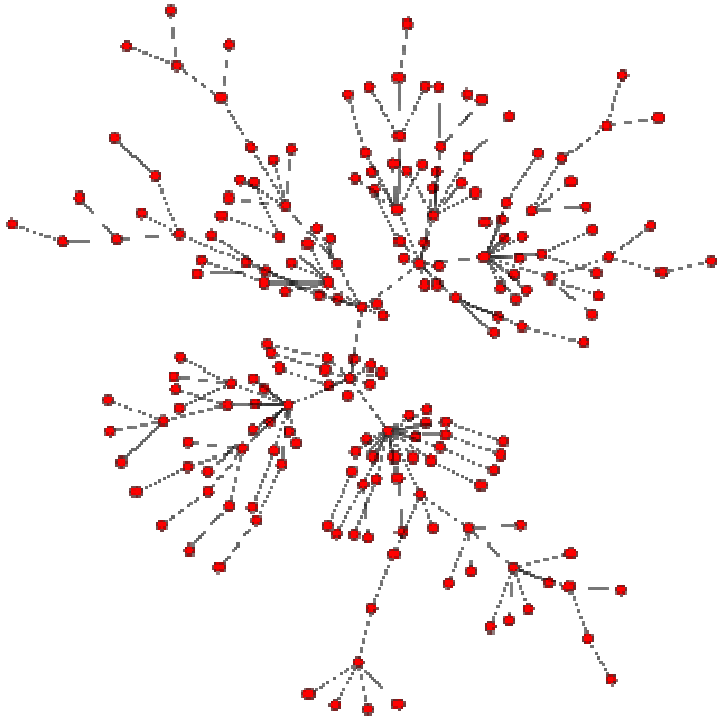
Distribúcia stupňov v reálnych sieťach - log škála (príklad DBpedia)



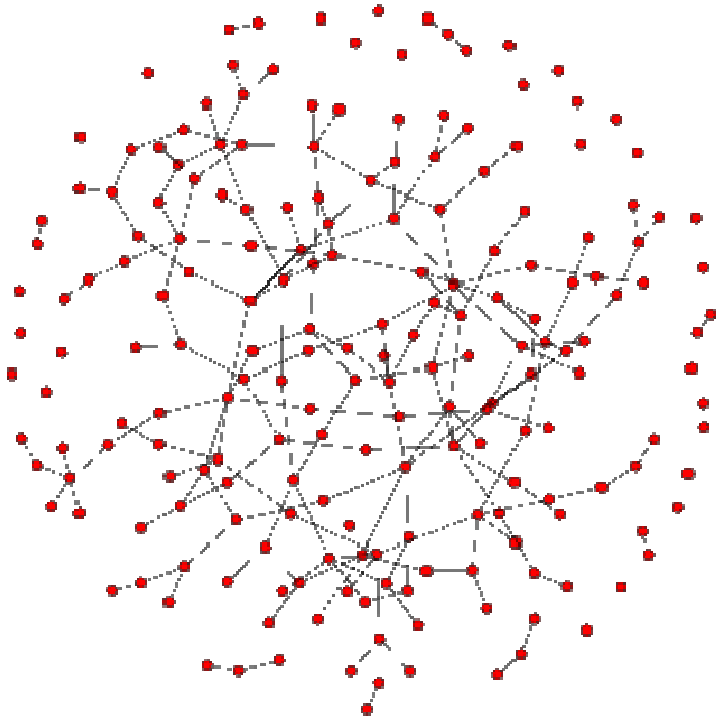
Zdroj: http://geza.kzoo.edu/bionet/html/scalefree.html

Náhodná sieť a sieť s mocninovou distribúciou

Sieť s mocninovou dist. stuňov



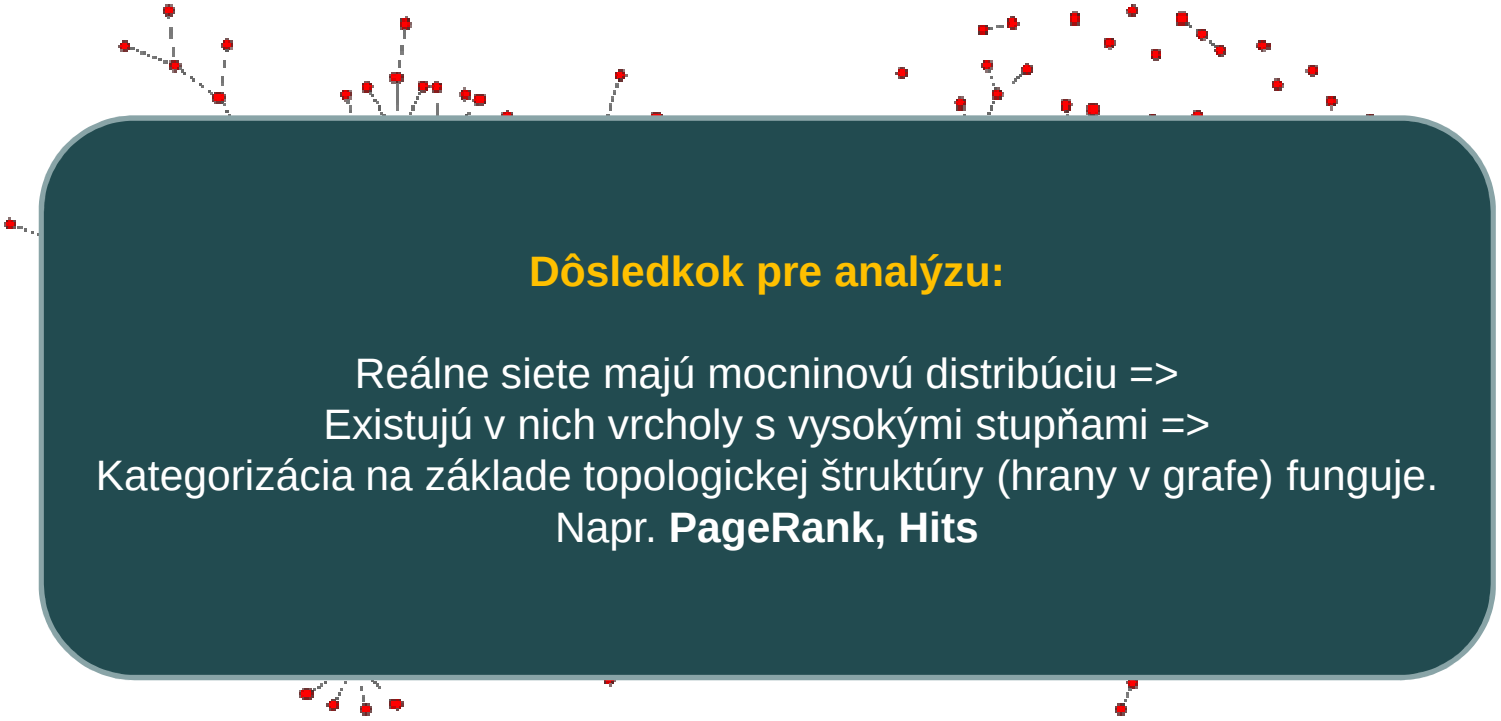
Sieť s binomickou dist. stuňov



Zdroj: http://geza.kzoo.edu/bionet/html/scalefree.html

Náhodná sieť a sieť s mocninovou distribúciou

Sieť s mocninovou dist. stuňov **Sieť s binomickou dist. stuňov**



Dôsledkok pre analýzu:

Reálne siete majú mocninovú distribúciu =>
Existujú v nich vrcholy s vysokými stupňami =>
Kategorizácia na základe topologickej štruktúry (hrany v grafe) funguje.
Napr. **PageRank, Hits**

Zdroj: http://geza.kzoo.edu/bionet/html/scalefree.html

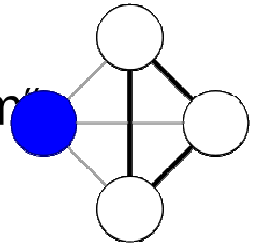
Siete malého sveta

- Objavené štúdiom sociálnych sietí v sociológii
- Milgramov experiment
  - **(60-te roky; úloha: doručenie listov cez sociálnu sieť od náhodných ľudí v Nebraske prímateľom v Chicagu; priemer 6 krokov pre)**
- Objav: sociálne site majú *krátku* dĺžku najkratších ciest medzi náhodne vybranými
- Potvrdené aj skúmaním počítačových sociálnych sietí
  - **Analýza MSM site – instant messaging**
  - **Priemerná dĺžka najkrašej siete medzi náhodne vybranými uzali: 6,6**
  - *[Jure Leskovec, Eric Horvitz: Planetary-scale views on a large instant-messaging network. WWW 2008]*
- Priemer site (aj priemerná dĺžka cesty medzi 2 vrcholmi) sa zmenšuje pri zväčšovaní siete
  - V protiklade k náhodným sieťam

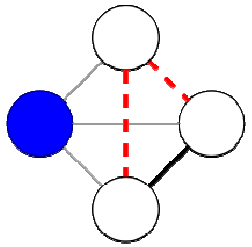
Zhlukovací koeficient

- Siete malého sveta často obsahujú kliky, alebo „skoro kliky“
- Efekt „moji priatelia v sociálnej sieti sú často priatelia navzájom“
- Matematicky to možno zachytiť pomocou zhlukovacieho koeficientu
- Lokálny zhlukovací koeficient:

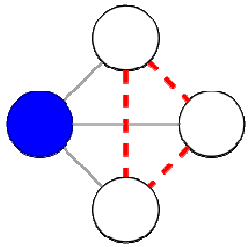
$$C_i = \frac{|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E.$$



c = 1



c = 1/3



c = 0

Zdroj: http://en.wikipedia.org/wiki/Clustering_coefficient

Vlastnosti reálnych sietí

- Mocninová distribúcia stupňov vrcholov
- Malá vzdialenosť medzi náhodnými uzlami v sieti (malý priemer grafu)
- Priemer grafu sa znižuje pri rozširovaní siete
- Vzor zhlukovania v sieti: vysoký zhlukovací koeficient

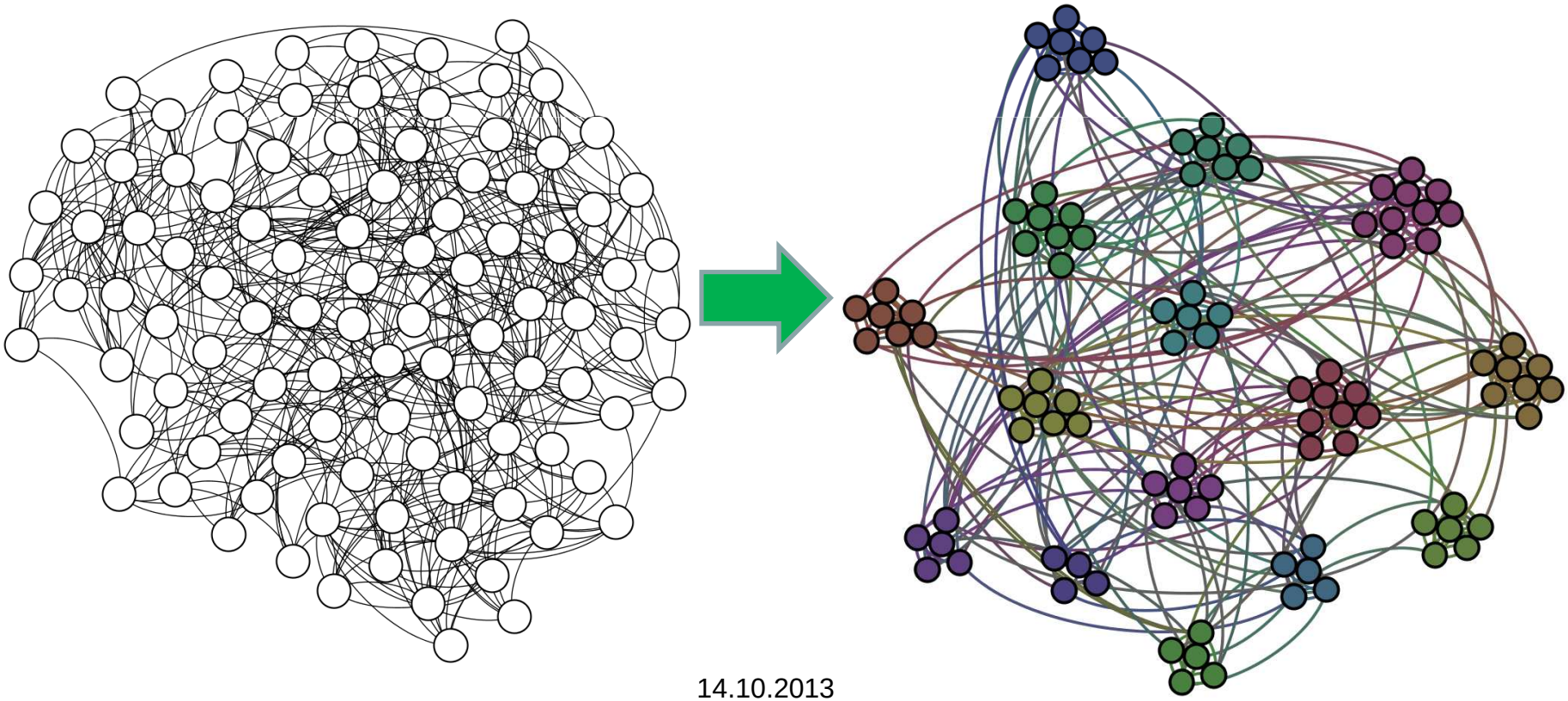
- Dôsledky:
  - **Algoritmy na spracovanie/dolovanie grafov fungujú vďaka týmto vlastnostiam**
  - **Rozličné typy sietí z reálneho sveta majú podobné vlastnosti**
  - **Algoritmy navrhnuté pre jeden konkrétny typ sietí (napr. soc. siete) budú pravdepodobne dávať zmysluplné výsledky aj na iných sieťach s podobnými matematickými vlastnosťami**
  - **Mocninová distribúcia stupňov – pri traversovaní grafu do šírky už pri nízkom počte hopov je nutné prejsť značnú časť siete**

Typické úlohy dolovania v grafoch

- Rekurzívne počítanie mier centralít:
  - **Odhadnúť dôležitosť vrchlov v topológii siete**
    - PageRank
    - HITS
- Detekcia komunít
  - **Identifikovať skupiny vrcholov silne prepojené medzi sebou a slabšie prepojené s ostatnými komunitami**
  - **Prekrývajúce sa komunity**
- Propagácia v sieťach
  - **Šírenie informácií v sieťach**
  - **Kaskádové správanie, propagácia výrusov**
- Klasifikácia objektov na základe liniek
- Predikcia vzniku liniek
- Objavovanie častých vzorov

Detekcia komunít

- Identifikácie zhlukov uzlov siete silne prepojených medzi sebou a menej silne prepojených so zvyškom siete
- Detekcia komunít v grafoch má za cieľ identifikovať moduly a ich prípadnú hierarchickú organizáciu



Detekcia komunít – definícia

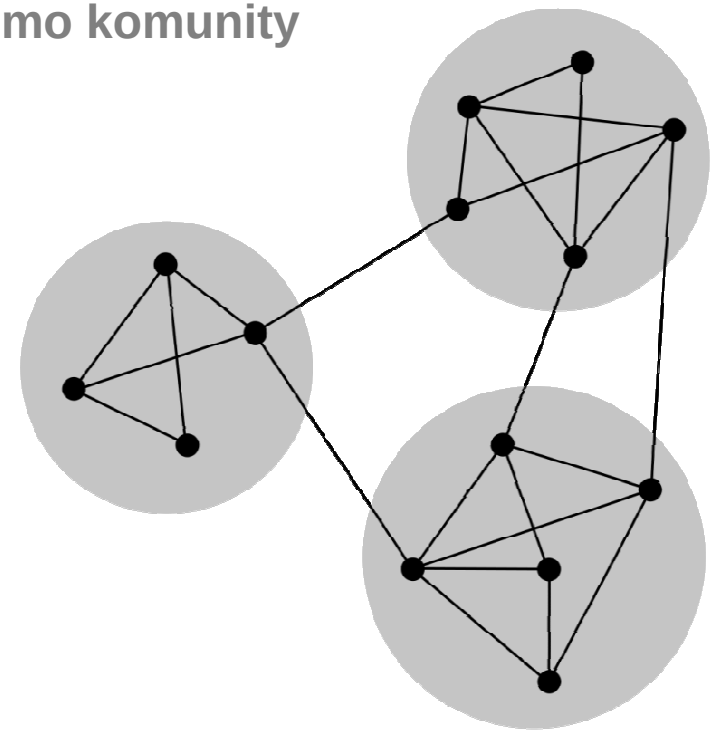
- Neexistuje konsenzus na formálnej definícii :
  - **problému detekcie komunít**
  - **komunitnej štruktúry grafu (community structue)**
  - **komunity ako takej**

- Najčastejšie používané definície komunity / zhluku grafu sú založené na počte hrán v rámci komunity (hustota) a počte hrán vedúcich mimo komunitu

- Definície komuitnej štruktúry sú často silne závislé na problémovej doméne a vlastnostiach analyzovaných grafov.

Detekcia komunít – definícia

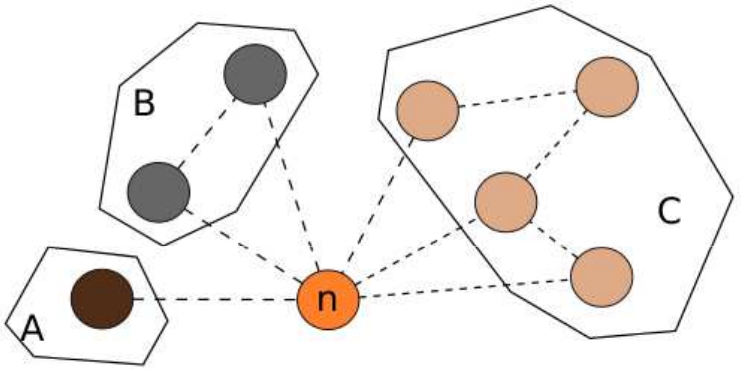
- Najčastejšie používaná definícia:
  - **Komunita je zhluk uzlov, kde počet vnútorných hrán v komunite je väčší ako počet vnokajších hrán – mimo komunity**



[M.E.J. Newman. The structure and function of complex networks. SIAM REVIEW 45, 2003.]

Detekcia komunít – definícia

- Alternatívna definícia:
  - **Komunita je zhluk uzlov, kde pre každý uzol platí, že počet hrán uzla v rámci komunity je väčší ako počet hrán uzla smerujúcich do inej komunity.**
  - **G = (V,E)**



$$aff(n, C) = \sum_{i \in C} w(e_{n,i}) : e_{n,i} \in E$$

$$\bigcup_{k \in \langle 1,\ldots,m \rangle} C_k = V \quad \text{and} \quad \bigcap_{k \in \langle 1,\ldots,m \rangle} C_k = \emptyset$$

$$\forall j \in C_k, aff(j, C_k) \geq max\{aff(j, C_l), C_l \in \gamma\}$$

- Y. Hu, H. Chen, P. Zhang, M. Li, Z. Di, and Y. Fan. Comparative definition of community and corresponding identifying algorithm. Phys. Rev. E, 78(2):026121, Aug 2008
- Marek Ciglan , Kjetil Nørvåg: Fast detection of size-constrained communities in large networks, proceedings of WISE'10, LNCS Volume 6488/2010

Detekcia komunít – definícia

- Dôsledky neexistencie konsenzu na formálnej definícií problému:
  - **Rôzni autori používajú rôzne definície komunít**
  - **Typické riešenie problému, 2 kroky:**
    - Špecifikácia metriky kvality komunitnej štruktúry
    - Algoritmická technika pre zhlukovanie grafu optimalizujúca danú metriku

- Veľké množstvo publikácií
  - **Prehľadová práca [Fortunato10]**
    - 457 referencií
    - Citovaná viac ako 1800 krát (podľa služby google scholar)
  - **Publikačný pretlak - Počet publikovaných algoritmov na detekciu komunít možno rátať v stovkách**
  - **Dopyt „community detection“ v google scholar - 8000+ výsledkov**
  - **Dopyt „community detection algorithm“ v google scholar - 1500+ výsledkov**

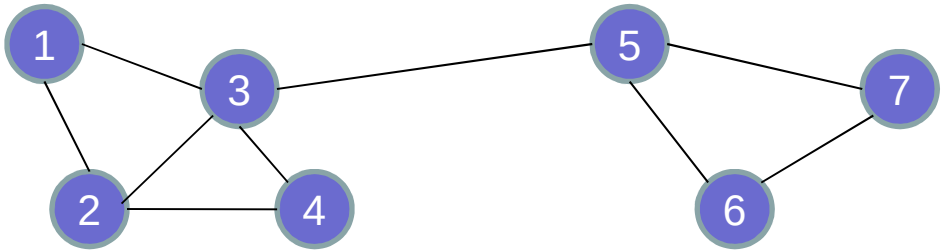
[Fortunato10] S. Fortunato (2010). "Community detection in graphs". Phys. Rep. 486 (3-5): 75–174. doi:10.1016/j.physrep.2009.11.002

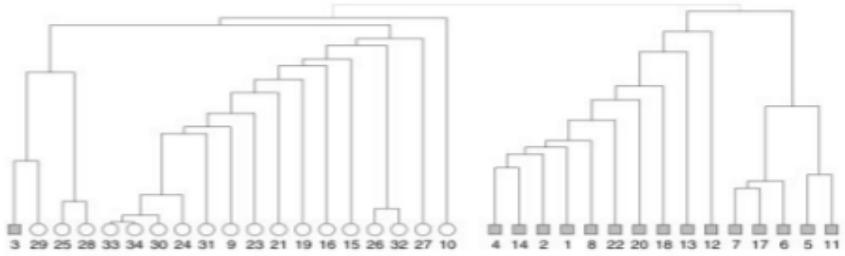
Triedy algoritmov na detekciu komunít

- Algoritmy založené na delení siete [Girvan, Newman]
- Algoritmy založené modularite
  - **Greedy aproximácie**
  - **Simlulované žíhanie**
  - **Extremal optimization**
- Algoritmy založené na spektrálnej analýze grafu
  - **Využíva spektrum grafu (vlastné čísla a vektory maticovej reprezentácie grafu)**
- Dynamické algoritmy
  - **Náhodné pochôdzky (random walk)**
  - **Spinové modely**
- Modely založené na štatistickej inferencii

Girvan-Newmanov algoritmus

- Založený na koncepte centrality spojitosti hrán (betweenness centrality)
- Centralita spojitosti pre e:
  - **Počet najkratších ciest prechádzjúcich hranou *e***
- Algoritmus:
  1. **Vypočítaj centralitu spojitosti pre všetky hrany grafu**
  2. **Hrana s najvyššou mierou centrality je odstránená**
  3. **Prepočítaj mieru centrality pre hrany ovplynené odobratím hrany**
  4. **Krok 2 a 3 je opakovaný do odobratia poslednej hrany**
- Výsledkom algoritmu je dendogram





Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad Sci. USA 99, 7821–7826 (2002)

Lokálne metriky kvality komunít

- Typický návrh algoritmu, 2 kroky:
  - **Špecifikácia *metriky* kvality komunitnej štruktúry**
  - **Algoritmická technika pre zhlukovanie grafu optimalizujúca danú metriku**
- Metriky zamerané na:
  - **Početnosť hrán v rámci zhluku**
  - **Početnosť hrán vedúcich mimo zhluku**
  - **Kombinácia oboch**
- Notácia:

$$G = (V, E)$$
$$C : C \subseteq V$$
$$n_c = |C|$$
$$m_c = |F| : F \subseteq E \wedge \forall (i,j) \in F : i,j \in C$$
$$o_c = |G| : g \subseteq E \wedge \forall (i,j) \in F : i \in C \wedge j \notin C$$
$$d_v = |K| : K \subseteq E \wedge \forall (i,j) \in K : i = v \wedge j \in V$$

Metriky – vnútorná konektivita

- Vnútorná hustota (Internal density)
  - **Vnútorná hustota komunity C**

$$f(C) = \frac{m_c}{(n_c \times (n_c - 1))/2}$$

- Priemerný stupeň (Average degree)
  - **Priemerný počet vnútorných hrán na uzol**

$$f(C) = \frac{2 \times m_c}{n_c}$$

- Priemerný počet trojuholníkov
  - **Priemerný počet vnútorných trojuholníkov na uzol**

$$f(C) = \frac{|\{(v : v \in C, \{u, w \in C \wedge (v,u) \in E \wedge (v,w) \in E \wedge (u,w) \in E\} \neq \emptyset\})|}{n_c}$$

Metriky – vnútorná konektivita

- Vnútorná hustota (Internal density)
  - **Vnútorná hustota komunity C**

$$f(C) = \frac{m_c}{(n_c \times (n_c - 1))/2}$$

[F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. PNAS, 101(9), 2004.]

- Priemerný stupeň (Average degree)
  - **Priemerný počet vnútorných hrán na uzol**

$$f(C) = \frac{2 \times m_c}{n_c}$$

- Priemerný počet trojuholníkov
  - **Priemerný počet vnútorných trojuholníkov na uzol**

[J. Yang and J. Leskovec. Defining and Evaluating Network Communities based on Ground-Truth. In: ICDM, 2012.]

$$f(C) = \frac{|\{(v : v \in C, \{u, w \in C \wedge (v,u) \in E \wedge (v,w) \in E \wedge (u,w) \in E\} \neq \emptyset\})|}{n_c}$$

Metriky – vonkajšia konektivita

- Expanzia (Expansion)
  - **Počet hrán mimo komunity na uzol**

$$f(C) = \frac{o_c}{n_c}$$

- Pomer rezu (cut ratio)
  - **Pomer existujúcich hrán idúcich mimo komunity k počtu možných hrán**

$$f(C) = \frac{o_c}{n_c \times (n \times n_c)}$$

Metriky – vonkajšia konektivita

- Expanzia (Expansion)
  - **Počet hrán mimo komunity na uzol**

$$f(C) = \frac{o_c}{n_c}$$

[F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. PNAS, 101(9), 2004.]

- Pomer rezu (Cut ratio)
  - **Pomer existujúcich hrán idúcich mimo komunity k počtu možných hrán**

$$f(C) = \frac{o_c}{n_c \times (n \times n_c)}$$

[Fortunato10] S. Fortunato (2010). "Community detection in graphs". Phys. Rep. 486 (3-5): 75–174. doi:10.1016/j.physrep.2009.11.002]

Metriky – kombinácie vnútornej a vonkajšej konektivity

- Vodivosť (Conductance)
  - **Pomer počtu hrán vedúcich mimo komunitu ku celkovému počtu hrán**

$$f(C) = \frac{o_c}{2m_c + o_2}$$

- Maximálny vonkajší stupeň (Maximum out degree fraction)
  - **Maximálny pomer von idúcich hrán z vrchola komunity k jeho stupňu**

$$f(C) = max_{v \in C} \frac{|\{(v,u) \in E \wedge u \notin C\}|}{d_v}$$

- Priemerný vonkajší stupeň (Average out degree fraction)
  - **Priemerný pomer von idúcich hrán z vrchola komunity k jeho stupňu**

$$f(C) = \frac{1}{n_c} \sum_{v \in C} \frac{|\{(v,u) \in E \wedge u \notin C\}|}{d_v}$$

Metriky – kombinácie vnútornej a vonkajšej konektivity

- Vodivosť (Conductance)
  - **Pomer počtu hrán vedúcich m … rán**

$$f(C) = \frac{o_c}{2m_c + o_2}$$

Fan Chung. Spectral Graph Theory. CBMS Lecture Notes 92, AMS Publications, 1997.

- Maximálny vonkajší stupeň (Maximum out degree fraction)
  - **Maximálny pomer von idúcich hrán z vrchola komunity k jeho stupňu**

$$f(C) = max_{v \in C} |\{(v,u) \in E \wedge u \notin C\}|$$

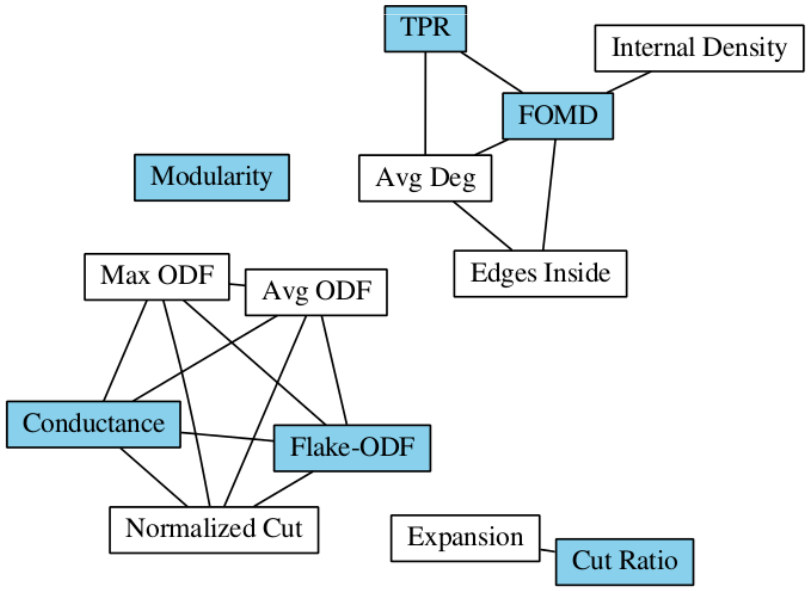
G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. In KDD '00, pp 150–160, 2000

- Priemerný vonkajší stupeň (Ave…
  - **Priemerný pomer von idúcich hrán z vrchola komunity k jeho stupňu**

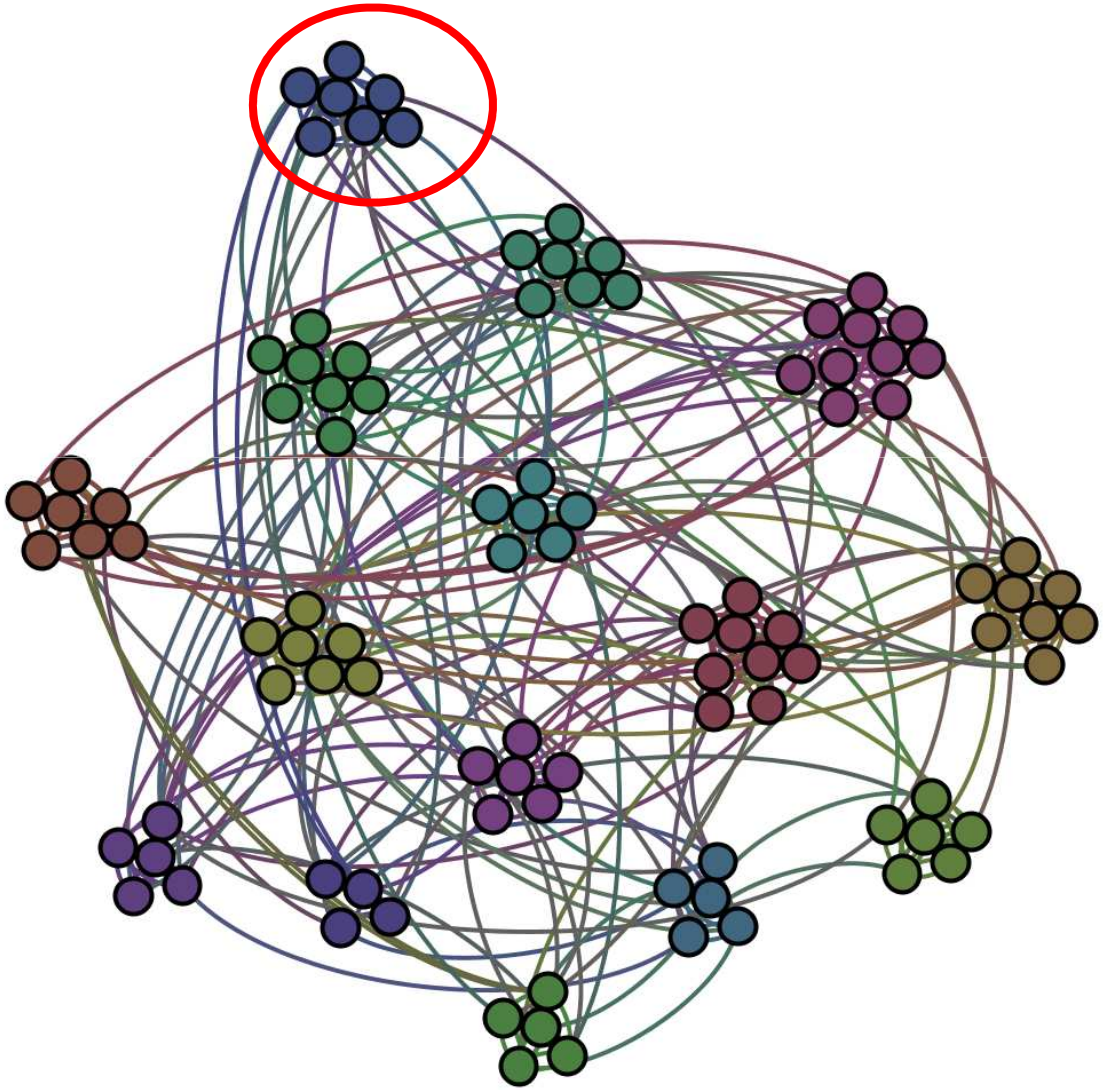
$$f(C) = \frac{1}{n_c} \sum_{v \in C} \frac{|\{(v,u) \in E \wedge u \notin C\}|}{d_v}$$

Metriky - porovnanie

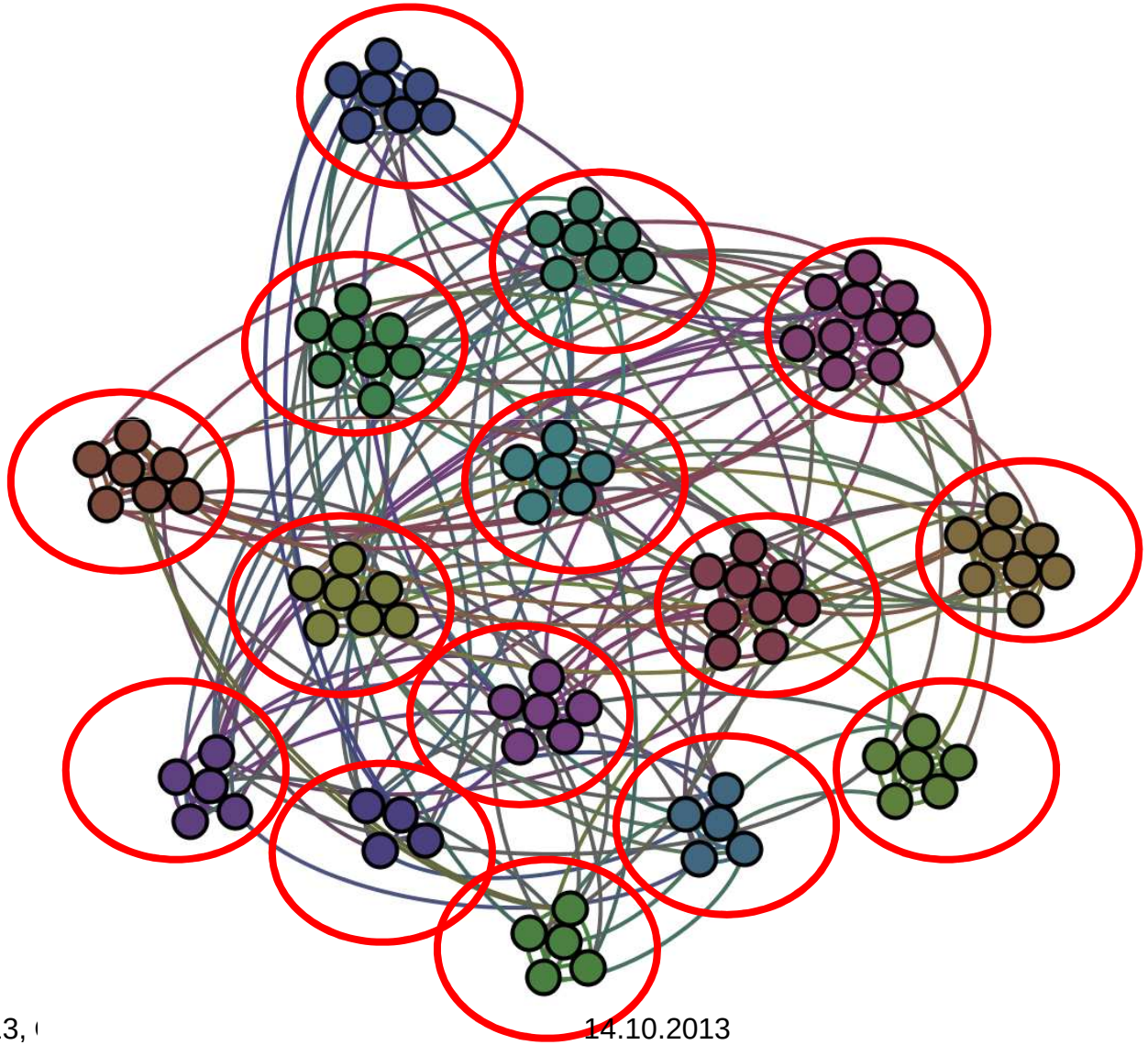
- Aký je vzťah medzi rozličnými metrikami?
- Zdroj: *[J. Yang and J. Leskovec. Defining and Evaluating Network Communities based on Ground-Truth. In: ICDM, 2012]*
- Autori použili reálne siete s explicitne definovanými komunitami
- Pre komunity vypočítali hodnoty jednotlivých metrík
- Porovnali koreláciu



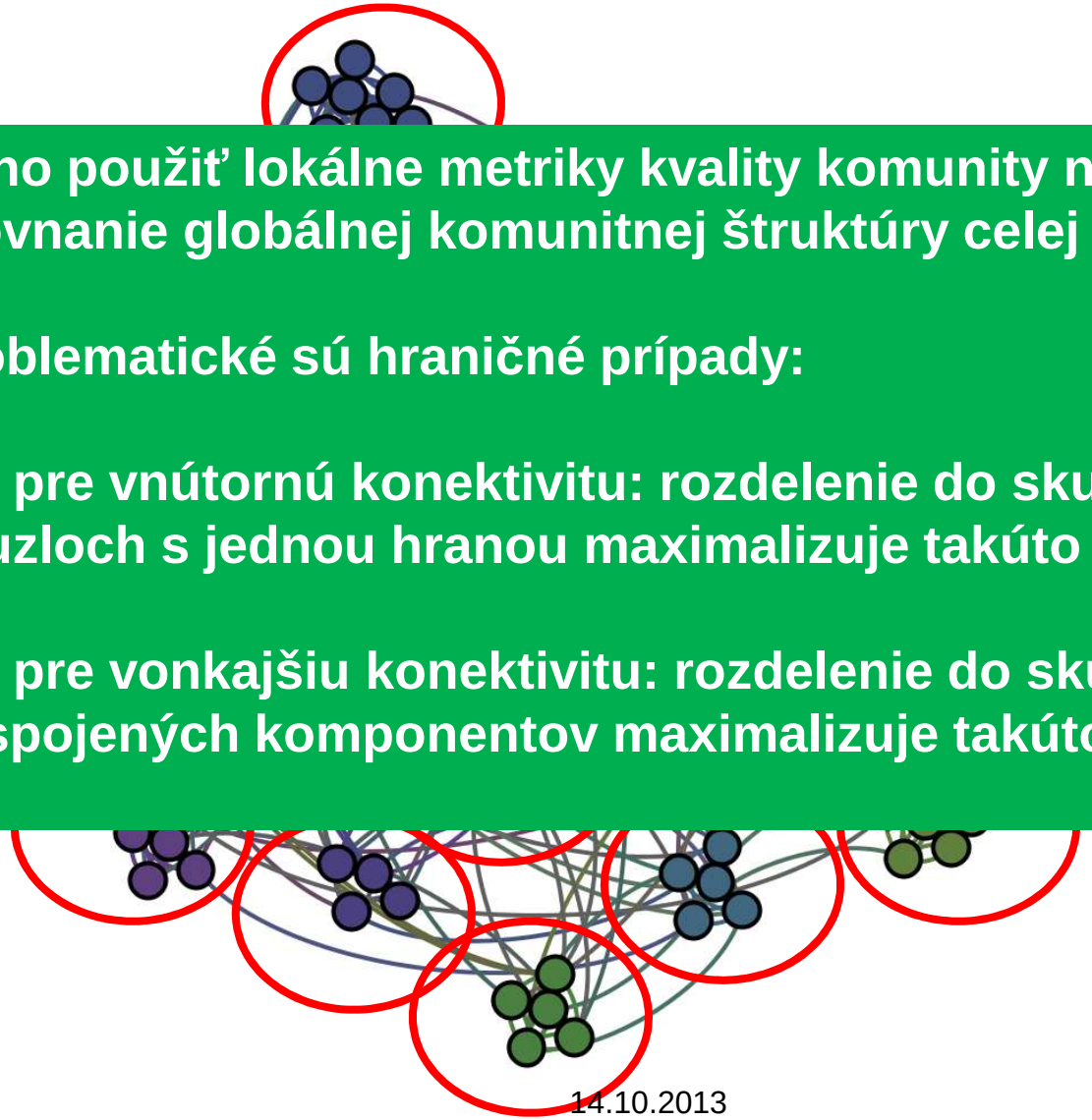
Metriky – globálne metriky rozdelenia siete



Metriky – globálne metriky rozdelenia siete



Metriky – globálne metriky rozdelenia siete



Možno použiť lokálne metriky kvality komunity na posúdenie / porovnanie globálnej komunitnej štruktúry celej siete?

- **Problematické sú hraničné prípady:**
  - **pre vnútornú konektivitu: rozdelenie do skupín po dvoch uzloch s jednou hranou maximalizuje takúto metriku**
  - **pre vonkajšiu konektivitu: rozdelenie do skupín na základe spojených komponentov maximalizuje takúto metriku**

Metriky - Modularita

- Zdoj: *[M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E 69(02), 2004.]*

- Modularita – metrika ohodnocujúca vhodnosť rozdelenia celej siete
- Myšlienka:
  - **Porovnať počet hrán ležiacich v zhlukoch s očakávaným počtom hrán v referenčnom modeli siete s rovnakým počtom vrcholov a hán**
  - **Referenčný model: obvykle náhodná sieť**
- Modularita
  - **Q = (početnosť hrán vo vnútri komunít) – (očakávaná početnosť hrán v rovnakom rozdelení do skupín v referenčnom modeli (v náhodnej sieti))**

Metriky - Modularita

- Modularita

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

- A_i,j – incidenčná matica
- k_i, k_j – stupne vrcholov i a j
- m – počet hrán grafu
- Delta(Ci, Cj) = 1 iff i a j ležia v rovnakom zhluku, ináč 0

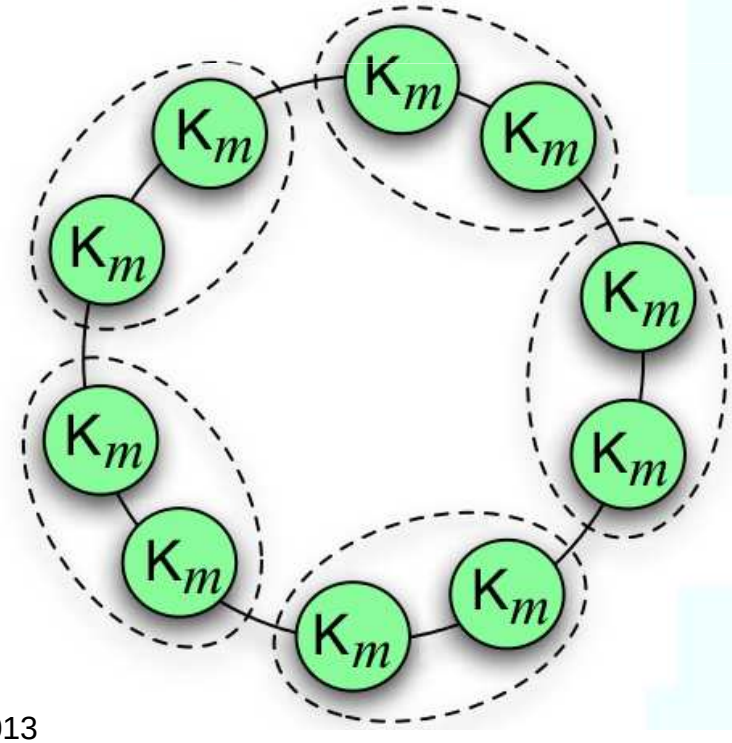
- Modularita vhodná ako
  - **Metrika kvality rozdelenia**
  - **Metrika na porovnanie rôznych rozdelení, algoritmov**

Modularita – limit rozlíšenia

- Práca: S. Fortunato and M. Barthelemy. Resolution limit in community detection. PNAS 104(1), 2007

- Problém modularity: **limit rozlíšenia**
- Optimalizácia modularity nerozozná malé komunity (relatívne, na počet hrán siete)

- Km – klik o m hranách
- m < sqrt(|E|)
  - **Maximálne skóre modularity odpovedá rozdeleniu s viacerými klikmi v komunitách**



Modularita – problém extrémnej degenerácie

- Práca: *[B. H. Good, Y.-A. de Montjoye and A. Clauset: The performance of modularity maximization in practical contexts. Physical Review E 81, 046106 (2010)]*

- Modularita vykazuje extrémnu degeneráciu - Exponenciálne množstvo rôznych rozdelení grafu do komunít nadobúda vysoké hodnoty modularity

Zhrnutie doterajšieho obsahu

- Neexistencia konseznu na formálnej špecifikácii problému
  - **Rôzne definície / intuitívne, neformálne definície**
  - **Veľké množstvo navrhnutých algoritmov**
- Metriky
  - **Lokálne – množstvo rôznych definícií**
  - **Lokálne metriky – zle transformovateľné na globálnu úroveň rozdelenia celej siete**
- Modularit
  - **Metrika ohodnotenia rozdelenia celej siete**
  - **Limit rozlíšenia – relatívne malé komunity nie sú odhalené**
  - **Problém degenerácie – veľa rozdelení s vysokým skóre modularity**

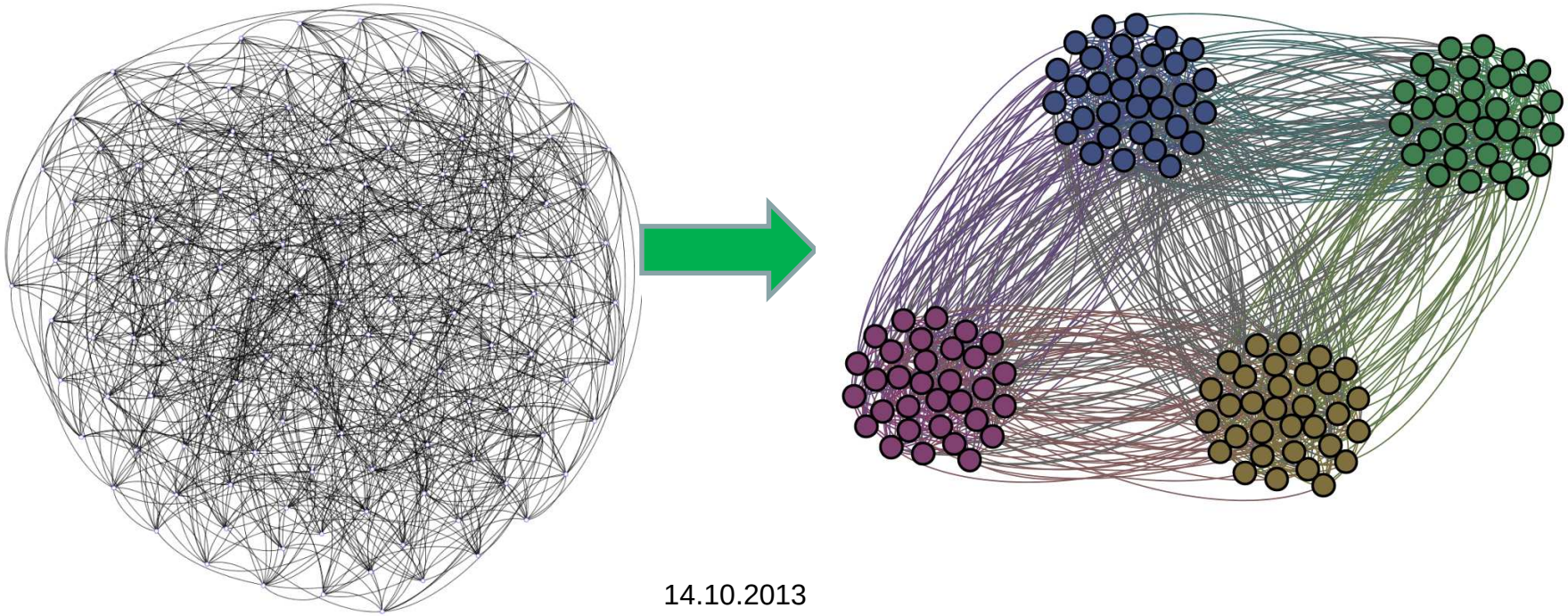
Skúšobné (benchmark) grafy

- Alternatívny prístup k vyhodnocovaniu presnosti algoritmov na detekciu komunít
- Majme grafy so známou štruktúrou komunít
- Výsledky zhlukovacieho algoritmu porovnáme s referenčným rozdelením

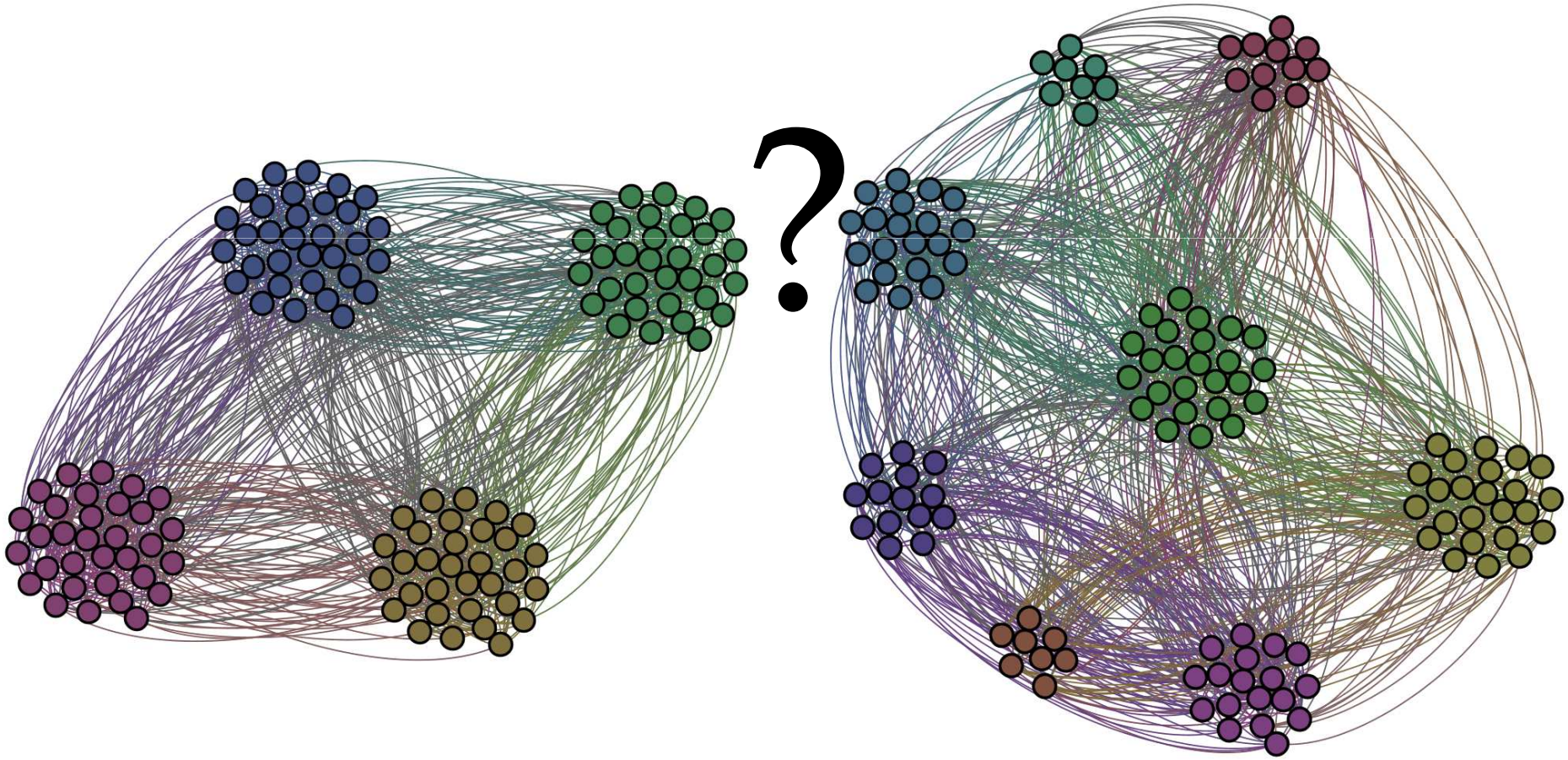
- Grafy možno generovať

Girvan-Newman benchmark grafy

- Práca: *[M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99, 7821 (2002).]*
- Benchmark grafy
  - **128 vrcholov**
  - **Rozdelných do 4-och komunít**
  - **Každá komunita: 32 uzlov**
  - **Priemerný stupeň uzla: 16**



Porovnanie detekovaného a referenčného rozdelenia



Normalized mutual information

- Práca: *[L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, Oct 2005]*

- Navrhnutá miera „Normalized mutual information“ (information theory)
- Výpočet založený na prienikovej matici *N* (confusion matrix)
  - **Riadky reprezentujú referenčné komunity**
  - **StÍpce reprezentujú detekované komunity**
  - **Prvky matice: N_ij = Počet prvkov referenčnej komunity i, ktoré sa nachádzajú v detekovanej komunite j**

$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{c_A} N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{c_B} N_{.j}\log\left(\frac{N_{.j}}{N}\right)}$$

Priemerná Jaccardova podobnosť

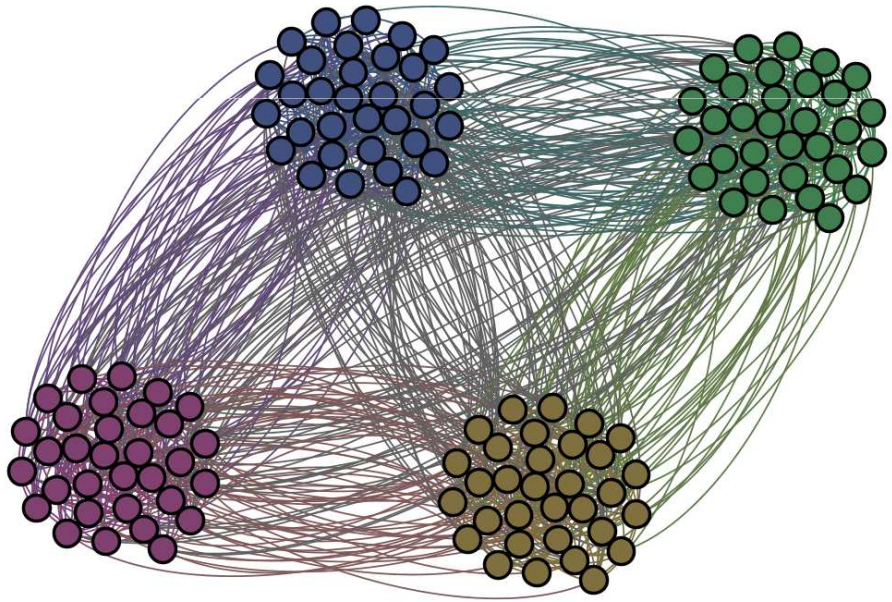
- Jaccardova podobnosť:
  - **štatistika určujúca podobnosť dvoch množín**

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Porovnáva dve množiny
- Použitie na porovnanie dvoch rozdelení do komunít:
  1. **Pre každú referenčnú komunitu**
  2. **nájdi detekovanú komunity s najväčším prienikom**
  3. **Vypočítaj Jaccardov index**
  4. **Výsledok: priemerný jaccardov index / alebo medián**

GN benchmark nets - kritika

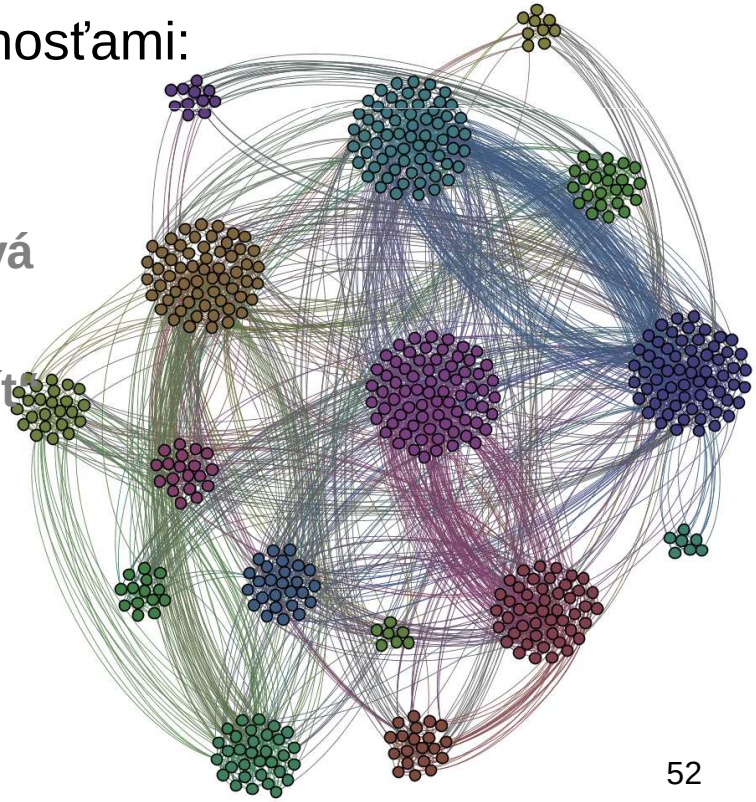
- GN benchmark siete:
  - **Nerealistické**
  - **Malé – 128 uzlov**
  - **Malý počet komunít**
  - **Rovnaká veľkosť komunít**
  - **Nerealistická distribúcia uzlov**



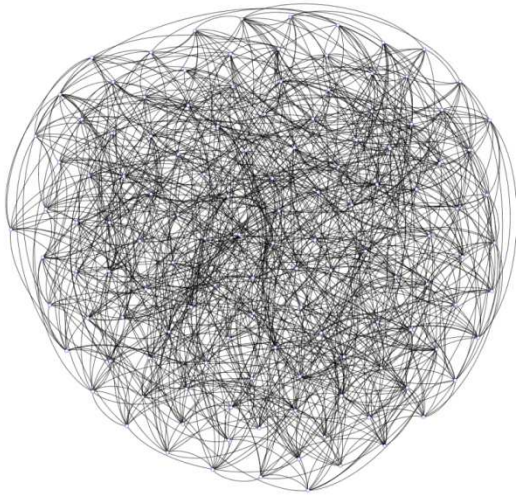
LFR benchmarks

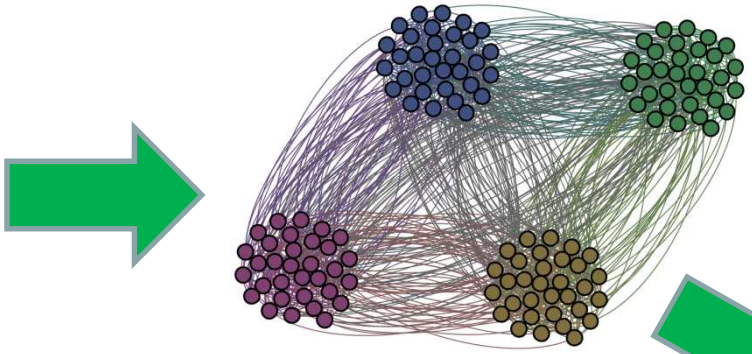
- Práca: *[A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E, 80(1), 2009.]*

- Snaha odstániť nedostatky GN-benchmark grafov
- Generovanie sietí s viac realistickými vlastnosťami:
  - **Založené na planted l-partition model**
  - **Distribúcia stupňov – mocninová**
  - **Distribúcia veľkostí komunít – mocninová**
  - **Definovateľný zhlukovací koeficient**
  - **Parametrizovateľná „viditeľnosť komunít"**



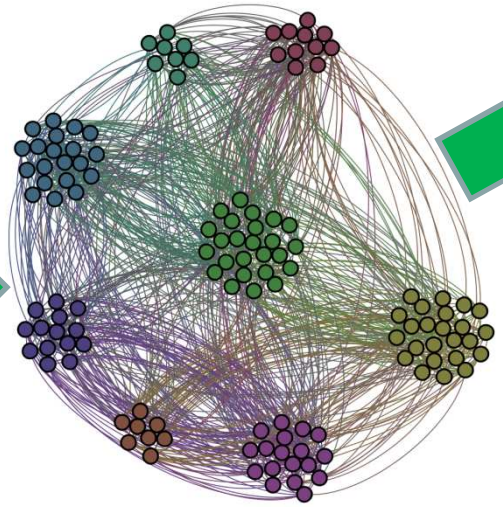
Rámec na vyhodnotenie algorimov





Metriky porovnania dvoch rozdelení
- NMI
- Priemerný Jaccardov index

Algoritmus na Detekciu komunít



Modifikácie základnej úlohy

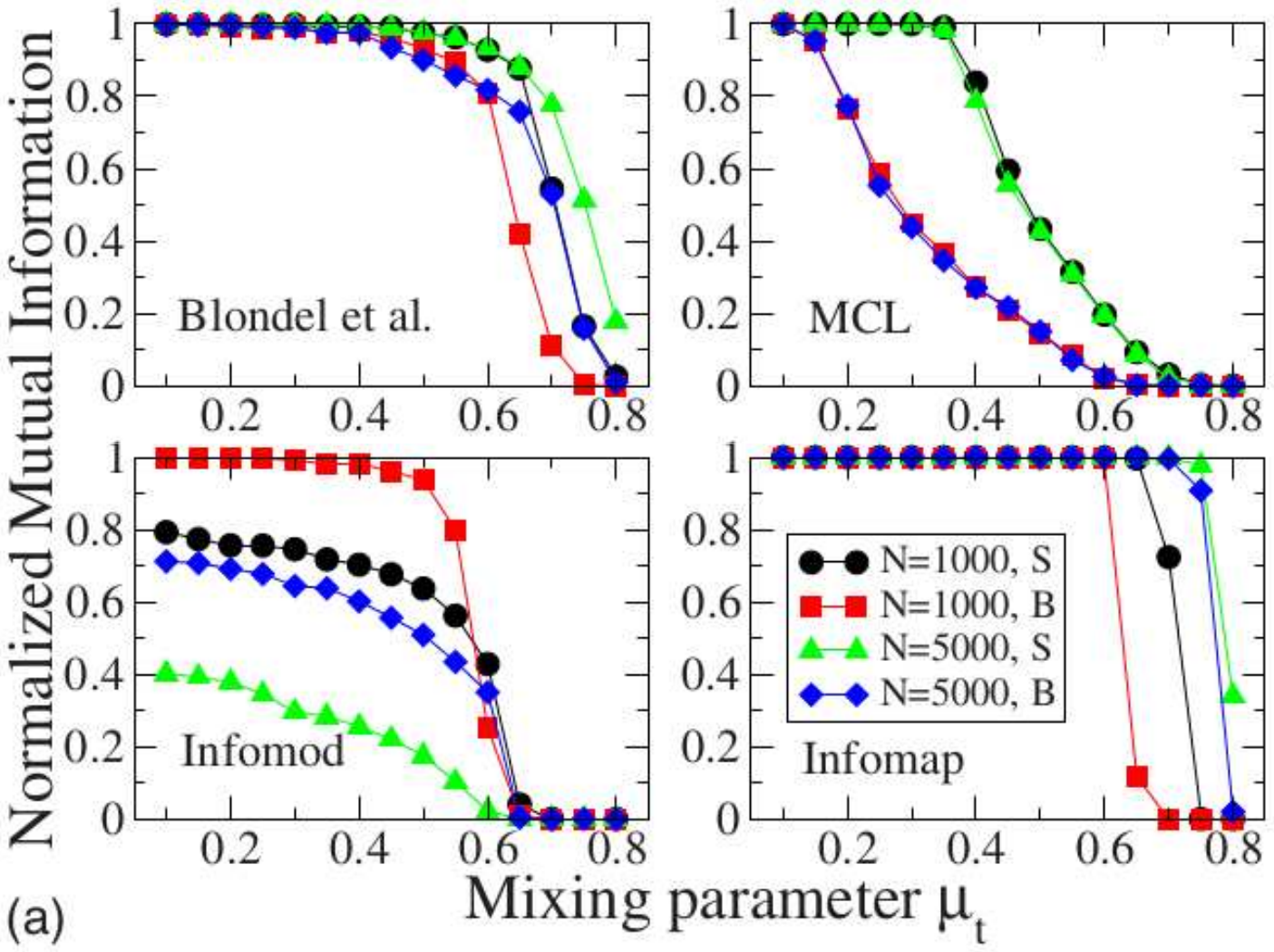
- Detekcia prekrývajúcich sa komunít
  - **Jeden vrchol patrí do viacerých komunít**
- Siete s atribútmi
  - **Vrcholy a/alebo hrany majú atribúty**
  - **Dodatočná informácia k topológii grafu**
- Detekcia komunity / komunít pre daný vrchol
- Hierarchická detekcia komunít

Porovnanie algoritmov na detekciu komunít

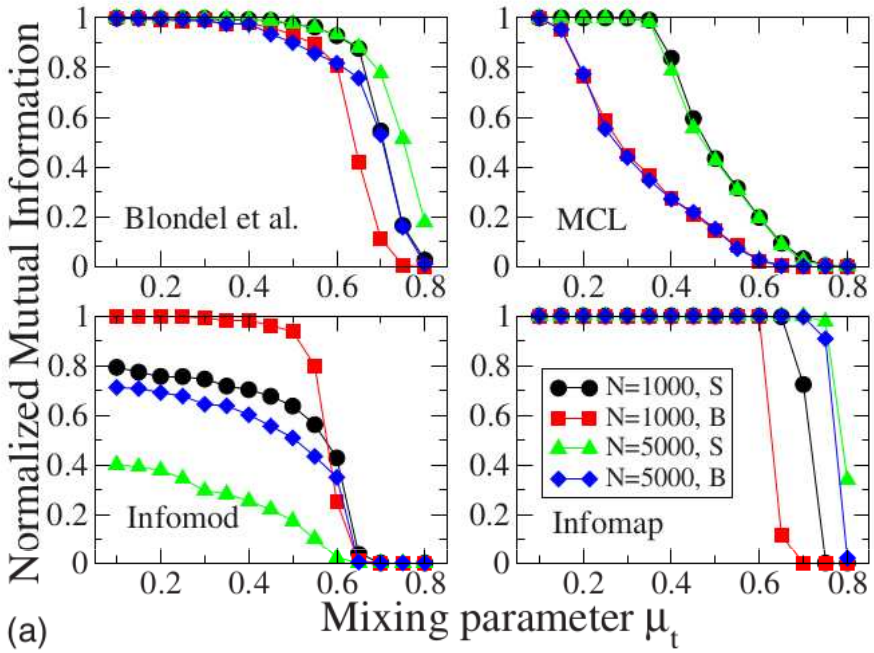
- Práca: *[A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. Phys. Rev. E, 80(5):056117, Nov 2009.]*

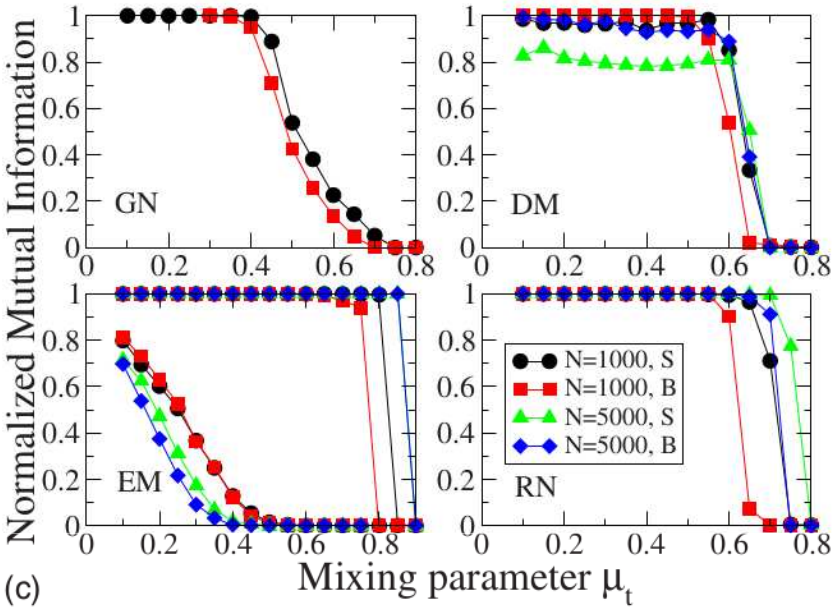
- Porovnanie populárnych algoritmov na detekciu komunít
- Použitie LFR bechmark grafov
- Porovnanie na rôznych nastaveniach generovaných sietí

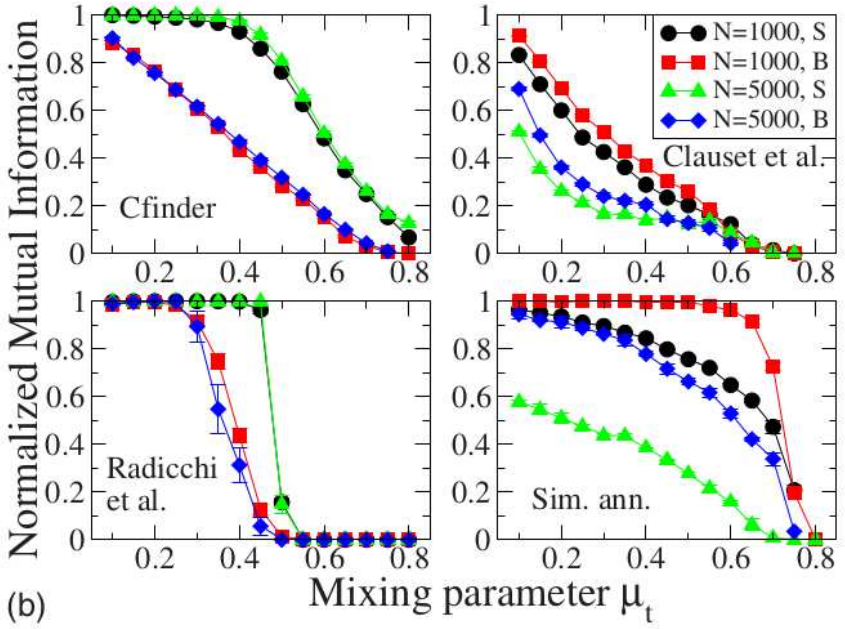
Porovnanie algoritmov na detekciu komunít



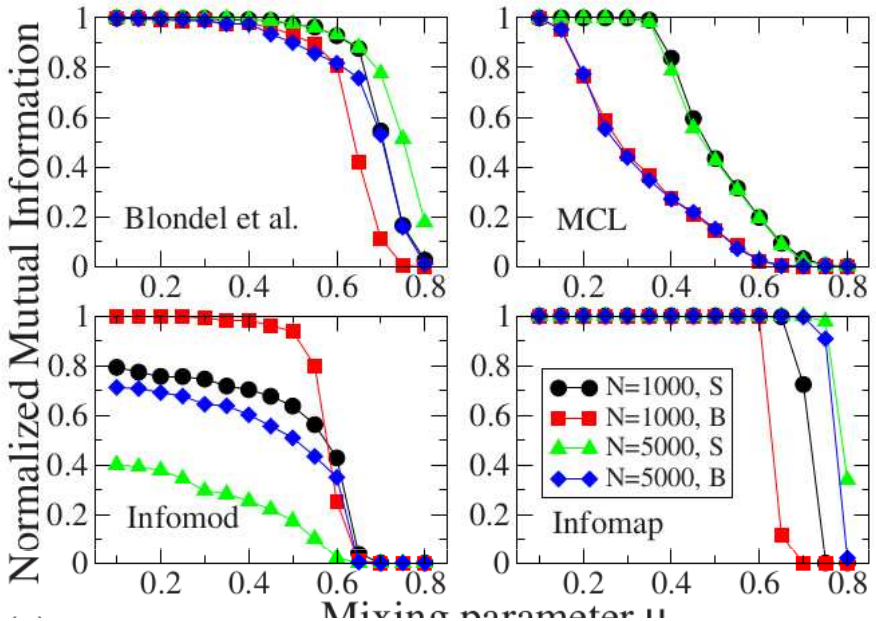
Porovnanie algoritmov na detekciu komunít

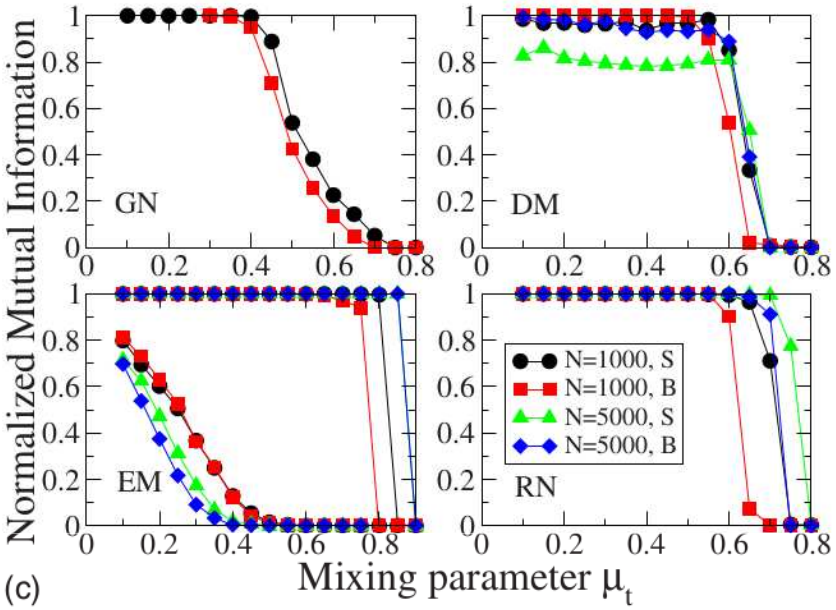


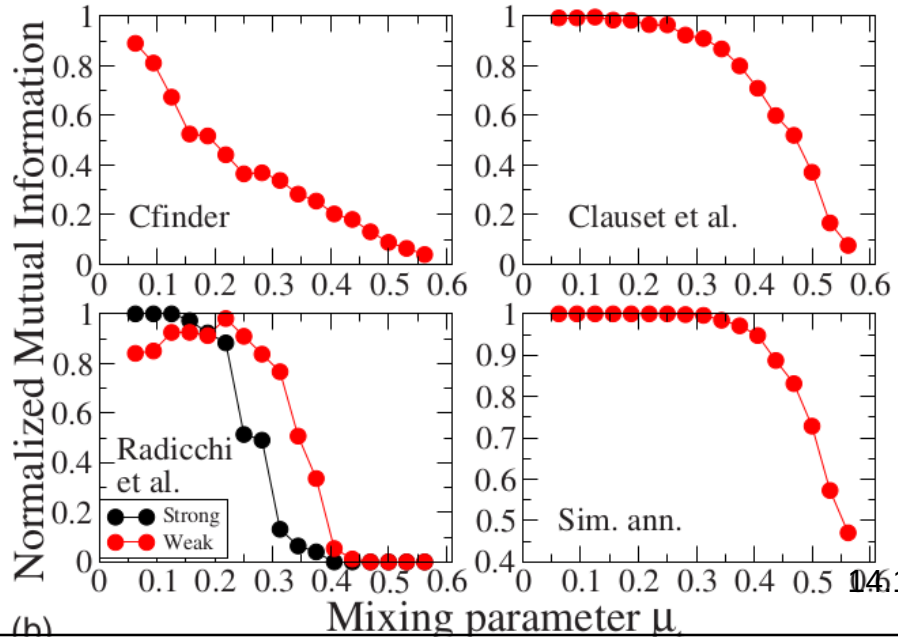


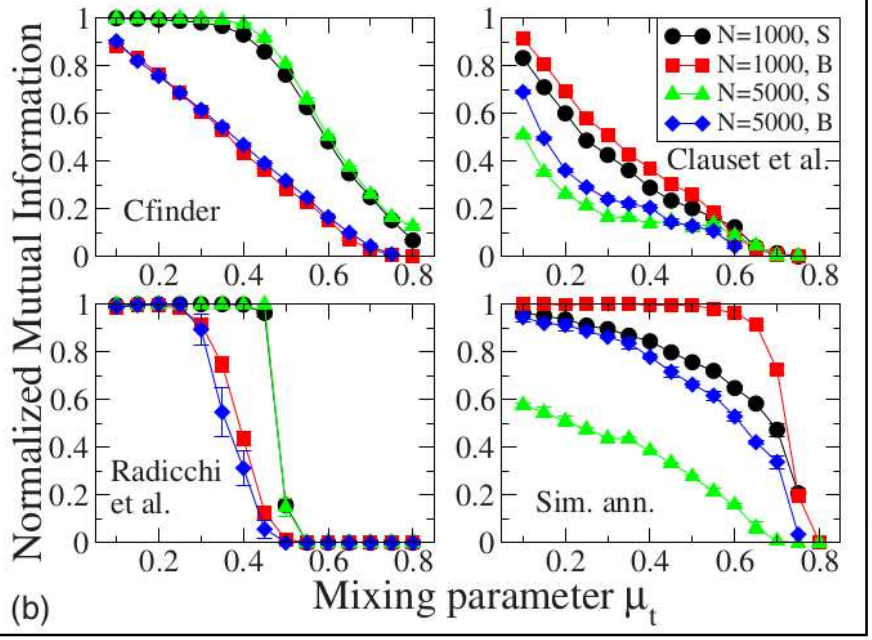


Porovnanie algoritmov na detekciu komunít









Detekcia komunít vo veľkých sieťach

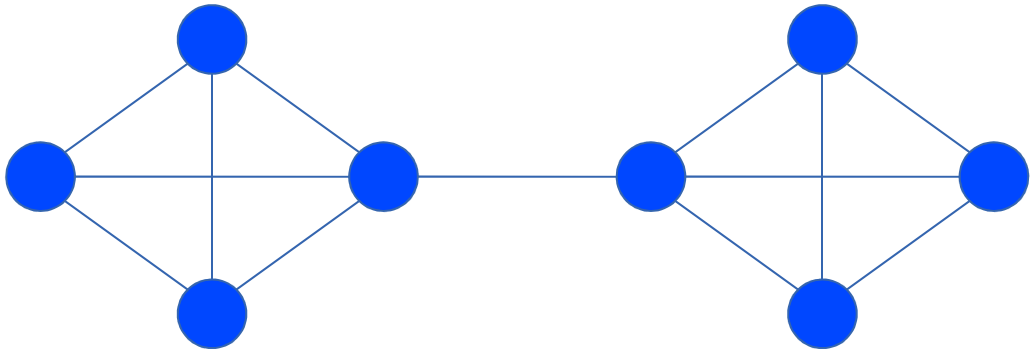
- Pamäťové obmedzenie
  - **Výpočty nad grafmi sú dátovo orientované**
  - **Prechod topológiou grafu**
  - **Práca s grafovými dátami je charakteristická náhodným prístupom k dátam**
  - **Práca s diskom vyžaduje veľa drahých I/O operácií**
  - **Štandardný prístup: celý graf v pamäti**

- Väčšina algoritmov ma detekciu má polynomiálnu časovú náročnosť
- Analýza veľkých sietí je problematická
- Príklad:
  - **Analýza grafu liniek Wikipédie**
  - **3+ miliónov vrcholov, 120+ miliónov hrán**
  - **Výpočet príliš dlhý**
- Pre veľké siete sú výhodné metódy s lineárnou časovou zložitosťou

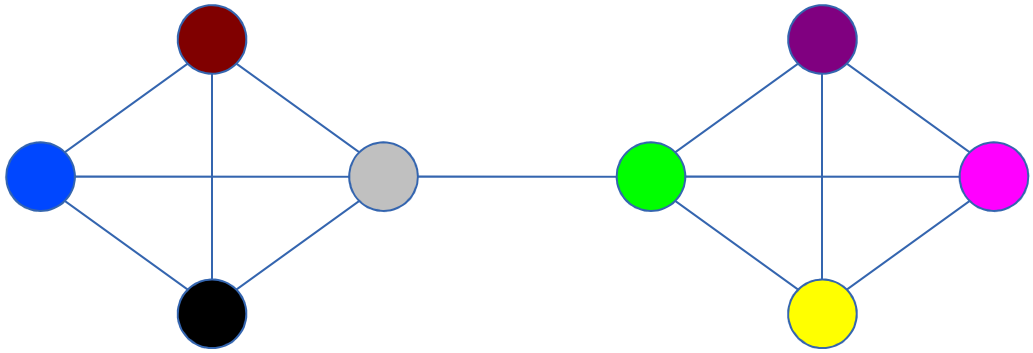
Algoritmus Propagácie značiek (Label Propagation)

- Greedy algoritmus s pseudo-lineárnou časovou náročnosťou
- Pseudo-lineárna časová náročnosť:
  - **Výpočet prebieha v iteráciách**
  - **Každá iterácia má lineárnu časovú zložitosť**
  - **Riešenie dobre konverguje po malom počte iterácií => Iterácií stačí relatívne malý počet**
- Algoritmus:
  - **Inicializácia: každý vrchol v separátnej komunite**
  - **Iterácie**
    - Prechod po uzloch v náhodnom poradí
    - Spracovávaný uzol priradíme do komunity, ku ktorej ho viaže najväčší počet hrán
    - V prípade viacerých komunít s najvyšším skóre, vyber príslušnosť náhodne

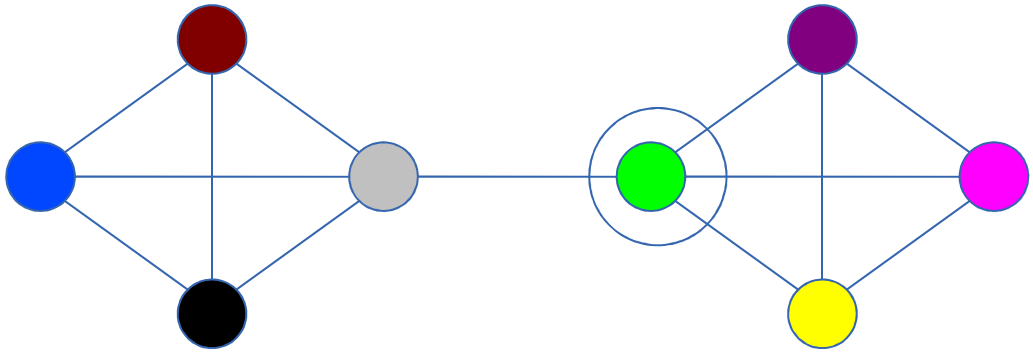
Propagácia značiek – Jednoduchý príklad



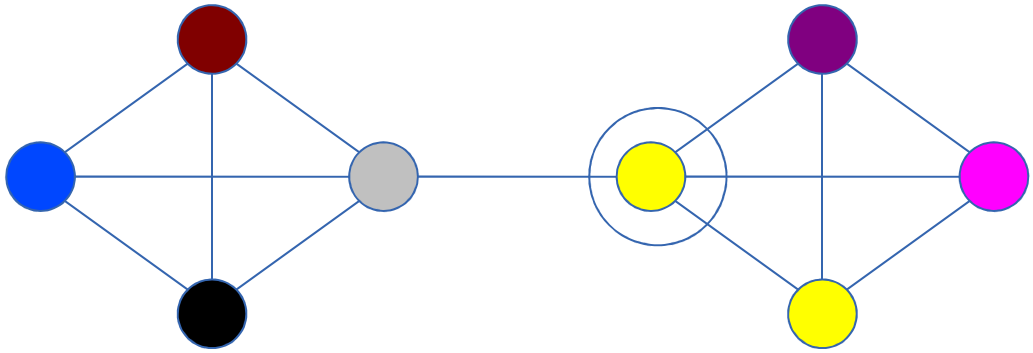
Propagácia značiek – Jednoduchý príklad



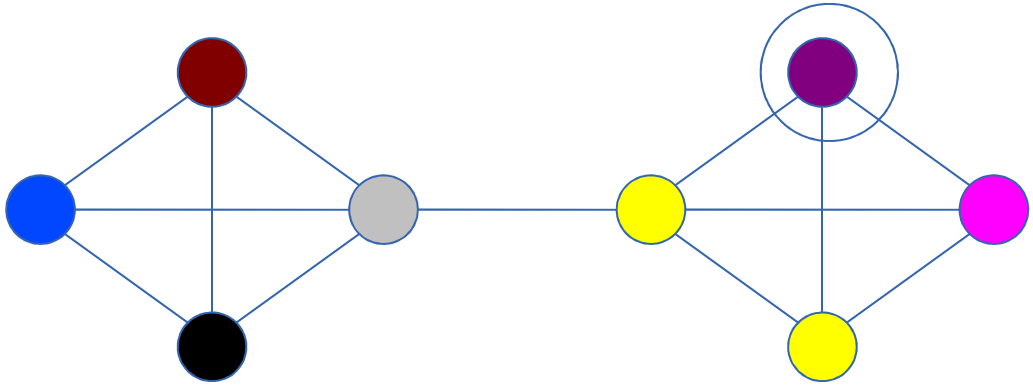
Propagácia značiek – Jednoduchý príklad



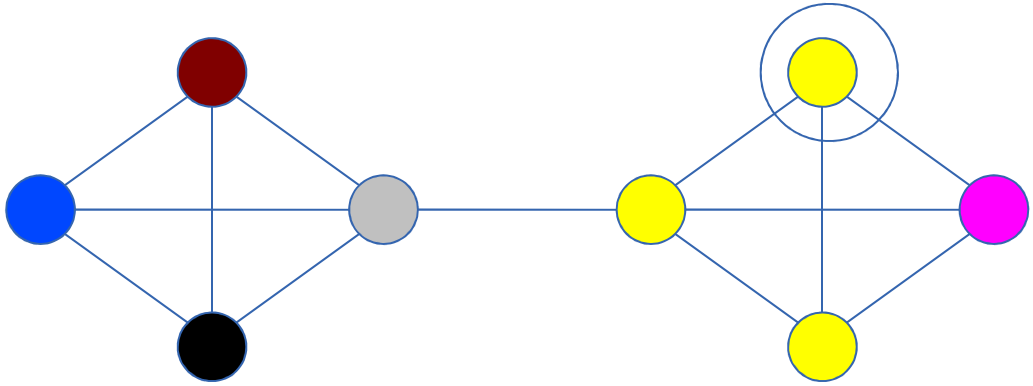
Propagácia značiek – Jednoduchý príklad



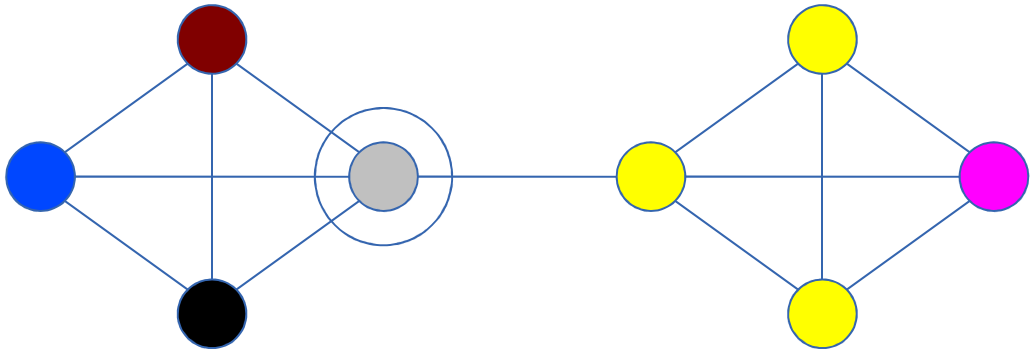
Propagácia značiek – Jednoduchý príklad



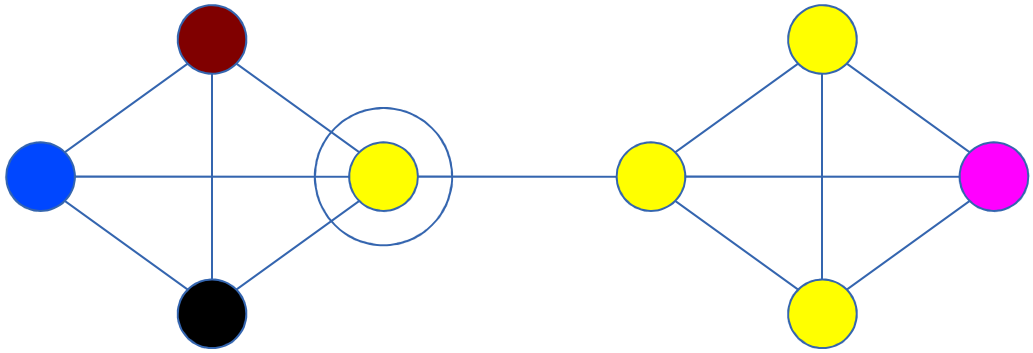
Propagácia značiek – Jednoduchý príklad



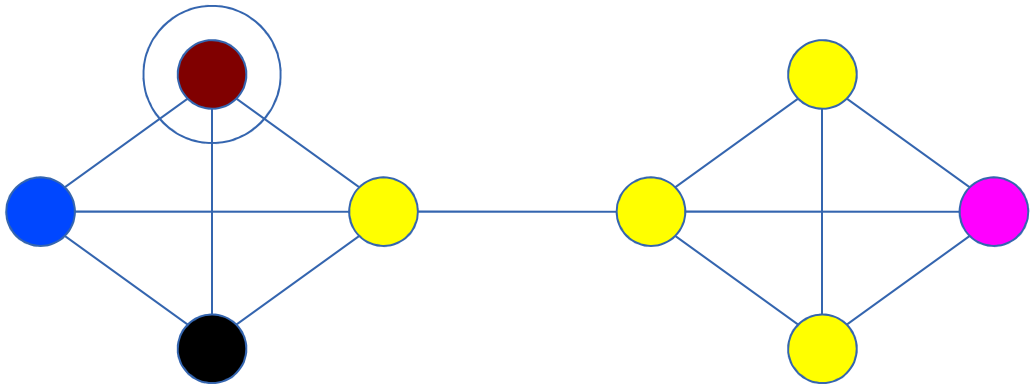
Propagácia značiek – Jednoduchý príklad



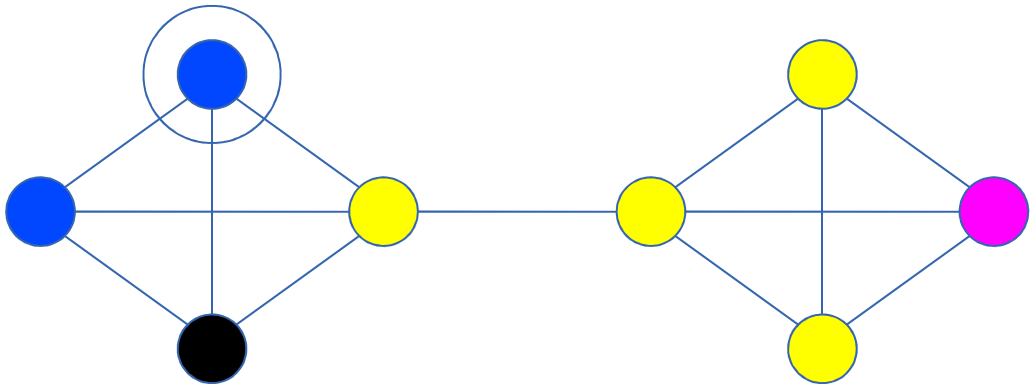
Propagácia značiek – Jednoduchý príklad



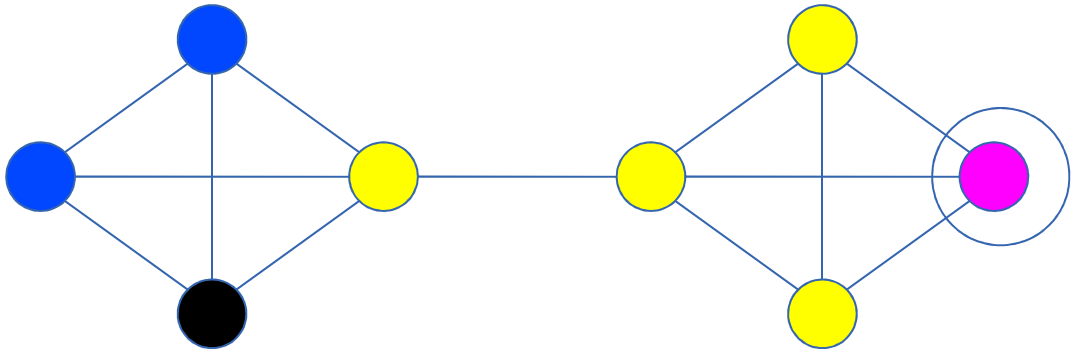
Propagácia značiek – Jednoduchý príklad



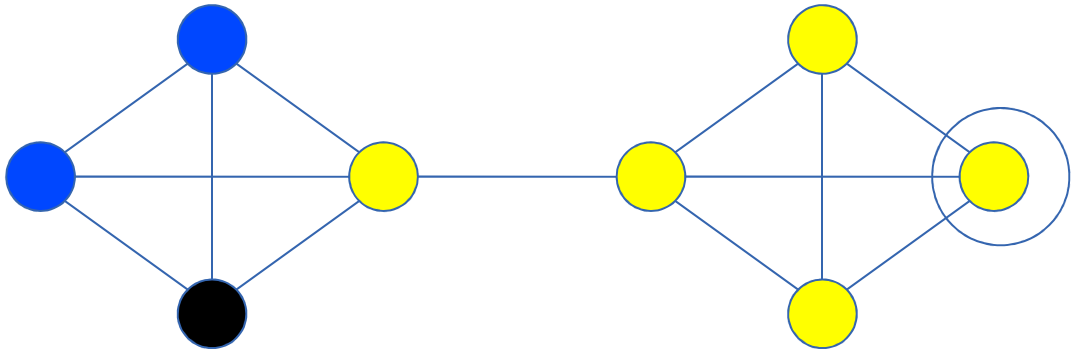
Propagácia značiek – Jednoduchý príklad



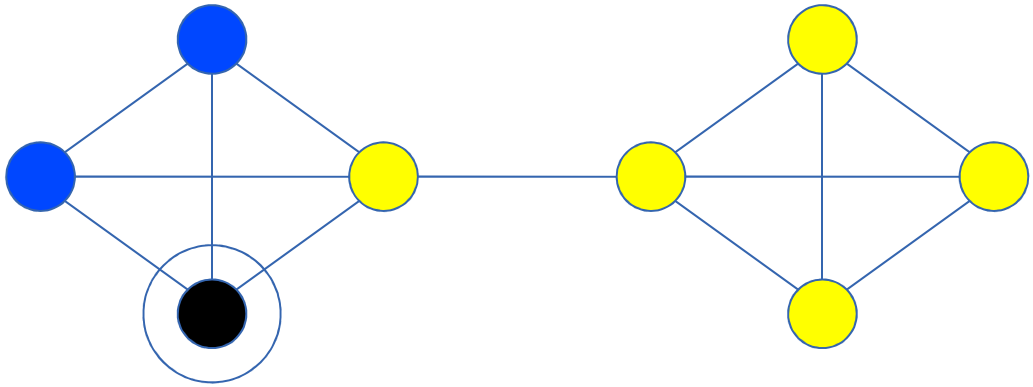
Propagácia značiek – Jednoduchý príklad



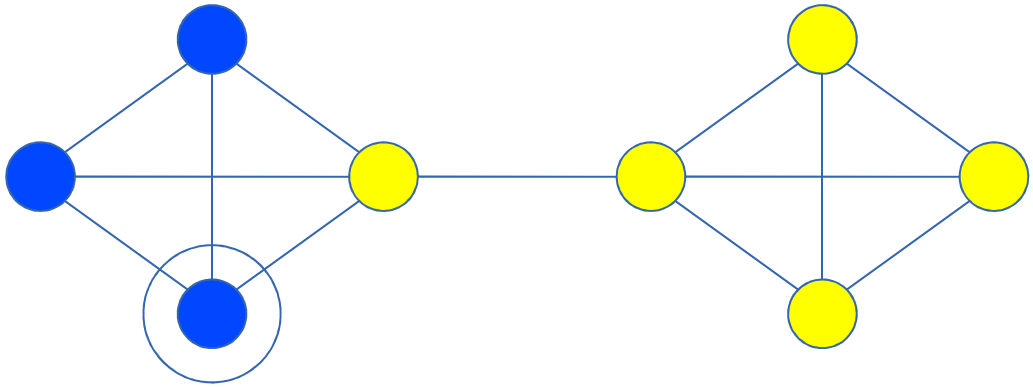
Propagácia značiek – Jednoduchý príklad



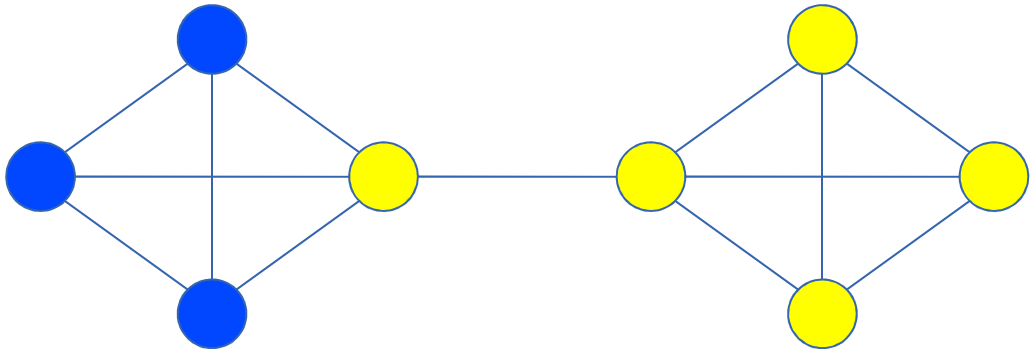
Propagácia značiek – Jednoduchý príklad



Propagácia značiek – Jednoduchý príklad

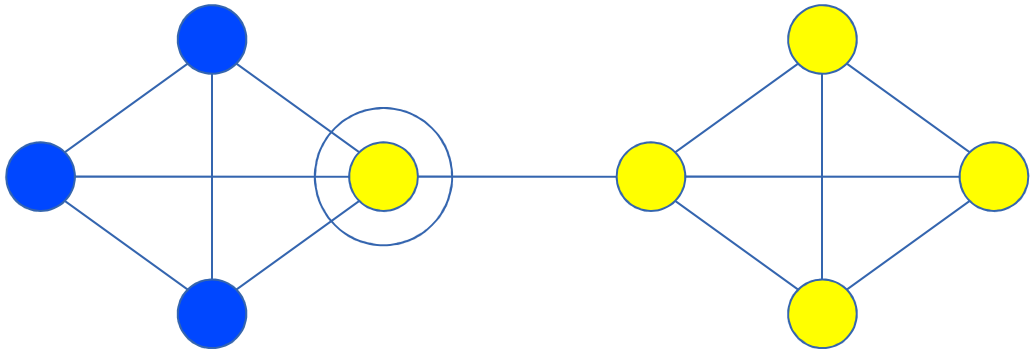


Propagácia značiek – Jednoduchý príklad

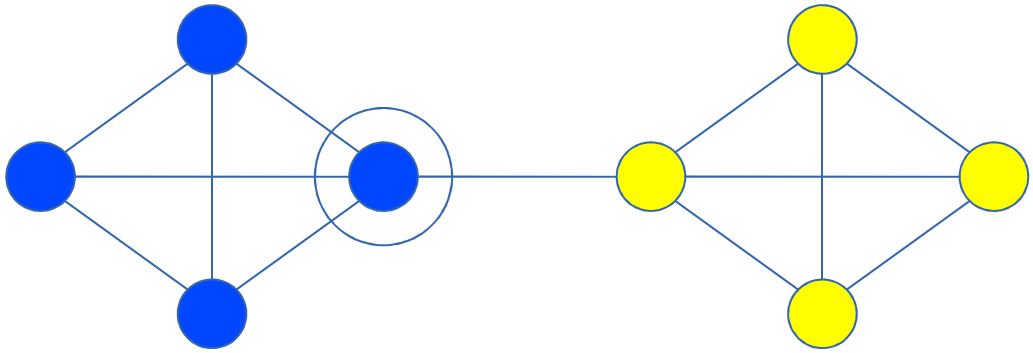


End of 1st iteration

Propagácia značiek – Jednoduchý príklad



Propagácia značiek – Jednoduchý príklad



Propagácia značiek

- Pre benchmark siete s nastavenou vysokou „viditeľnosti komunít“ má dobré výsledky
- Pri menej jasných konfiguráciách má tendenciu zahrnúť veľkú časť vrcholov do jednej komunity

Modifikácie algoritmu propagácie značiek

- Základný algoritmus:
  - **Inicializácia: každý vrchol v separátnej komunite**
  - **Iterácie**
    - Prechod po uzloch v náhodnom poradí
    - Spracovávaný uzol priradíme do komunity, pre ktorá má najvyššie hodnotu **funkcie nárastu**
    - V prípade viacerých komunít s najvyšším skóre, vyber príslušnosť náhodne

Louvain metóda

- Greedy algoritmus optimalizujúci modularitu
- Založený na princípe propagácie značiek
- Zachováva pseudo-lineárnu časovú závislosť
- Funkcia nárastu:

$$\Delta Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

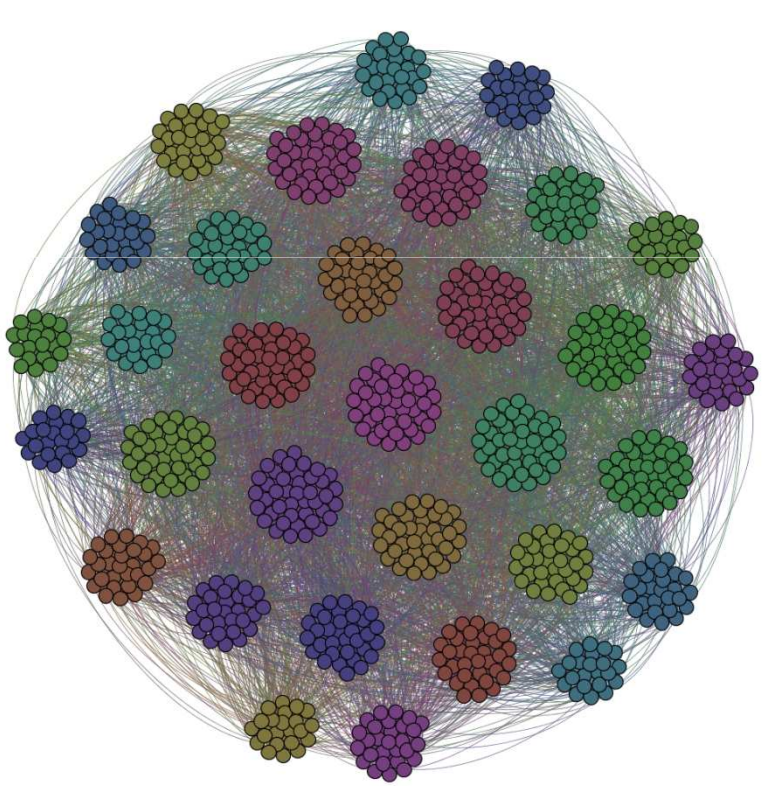
- SUM_in – suma váh liniek v komunite C
- SUM_tot - suma váh liniek hrán incidentých s uzlami komunity C
- k_i – váha liniek incidentých s uzlom k
- k_i,in – váha liniek incidentých s uzlom k v C
- m - váha všetkých liniek v sieti

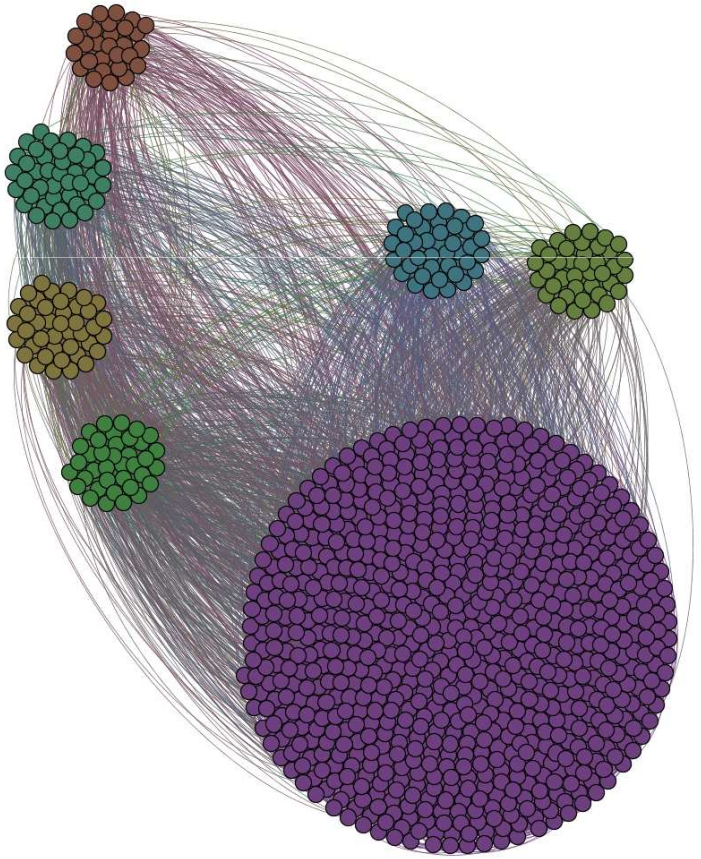
[Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre: Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008]

Cvičenie: Greedy detekcia komunít na Wikipédii

- Detekcia komunít v link grafe Wikipédie
  - **3.1M nodes**
  - **91M edges**
- Label Propagation algoritmus:
  - **Najväčšia kominta: 2.96M uzlov**
- Louvain metóda:
  - **Najväčšia kominta: 400K uzlov**
  - **20 najväčších komunít : 95% siete**

Cvičenie: Greedy detekcia komunít na Wikipédii





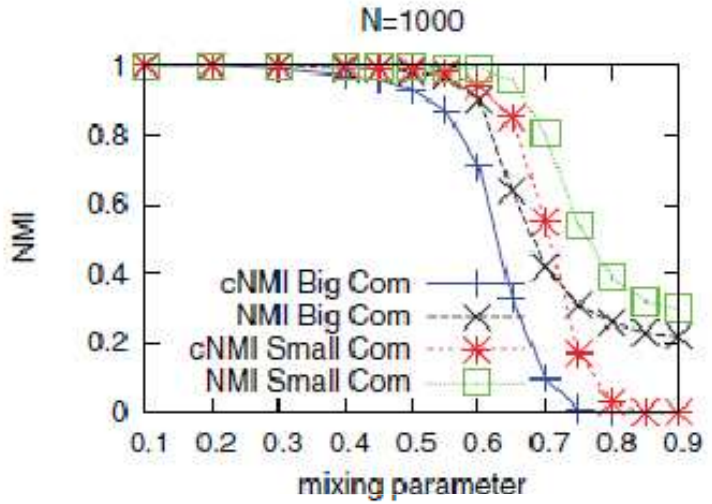
SCCD (Size Constrained Community Detection)

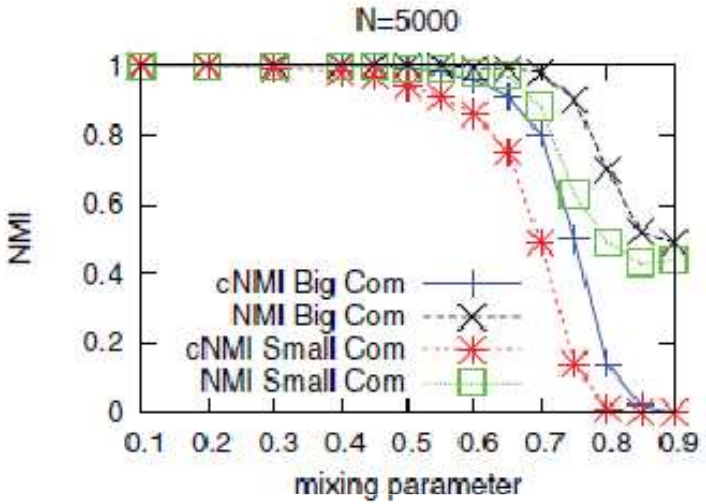
- Greedy algoritmus založený na princípe propagácie značiek
- Zachováva pseudo-lineárnu časovú závislosť
- Motivácia:
  - **Zabrániť vzniku obrovských zhlukov**
  - **Umožniť parametrizáciu maximálnej veľkosti komunity**
- Funkcia nárastu:

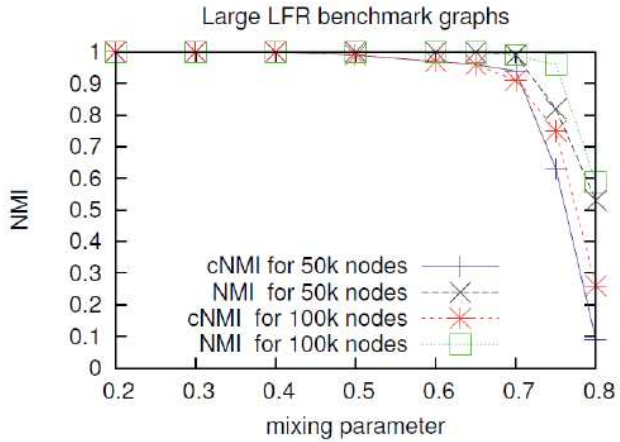
$$seed\_gain(n, C) = aff(n, C) \times log\left(\frac{UpperLimit}{|C|}\right)$$

[Marek Ciglan , Kjetil Nørvåg: Fast detection of size-constrained communities in large networks, proceedings of WISE'10, LNCS Volume 6488/2010, Springer-Verlag]

SCCD (Size Constrained Community Detection)







[Marek Ciglan , Kjetil Nørvåg: Fast detection of size-constrained communities in large networks, proceedings of WISE'10, LNCS Volume 6488/2010, Springer-Verlag]

Siete reálneho sveta s explicitnými komunitami

- Práca: *[J. Yang and J. Leskovec. Defining and Evaluating Network Communities based on Ground-Truth. In: ICDM, 2012]*

- Autori použili reálne siete s explicitne definovanými komunitami
- Siete z rôznych zdrojov:
  - **LiveJournal, Orkut, Youtube**
    - užívateľmi definované skupiny budeme pokladať za komunity
  - **DBLP**
    - Graf ko-autorstva publikácií
    - Autori publikujúci na rovnakých konferenciách a spojený v komponente => komunita
  - **Amazon**
    - Sieť s hranami indikujúcimi časté spolu-kupovanie dvoch produktov
    - Komunity definované kategóriami
- Filtrovanie explicitne definovaných komunít na základe lokálnych metrík kvality komunít

Detekcia komunít v rozsiahlych sieťach reálneho sveta

- Práca: *[Marek Ciglan, Michal Laclavík and Kjetil Nørvåg: On Community Detection in Real-World Networks and the Importance of Degree Assortativity, 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2013]*

- Použili sme rýchle algoritmy na detekciu komunít na niekoľkých rozsiahlych sieťach reálneho sveta so známymi explicitnými komunitami
  - **LiveJournal, Orkut, Youtube, DBLP, Amazon, DBPedia**
  - **DBPedia: znalostná báza derivovaná z Wikipédie; explicitné komunity sú kategórie Wikipédie**
  - **Použité algoritmy: Label propagation, Louvain method, SCCD**
- Pre niektoré siete algoritmy doručili dobrú aproximáciu explicitne definovaných zhlukov, pre časť sietí nie

Detekcia komunít v rozsiahlych sieťach reálneho sveta

- Porovnanie detekovaného rozdelenia siete a explicitných komunít

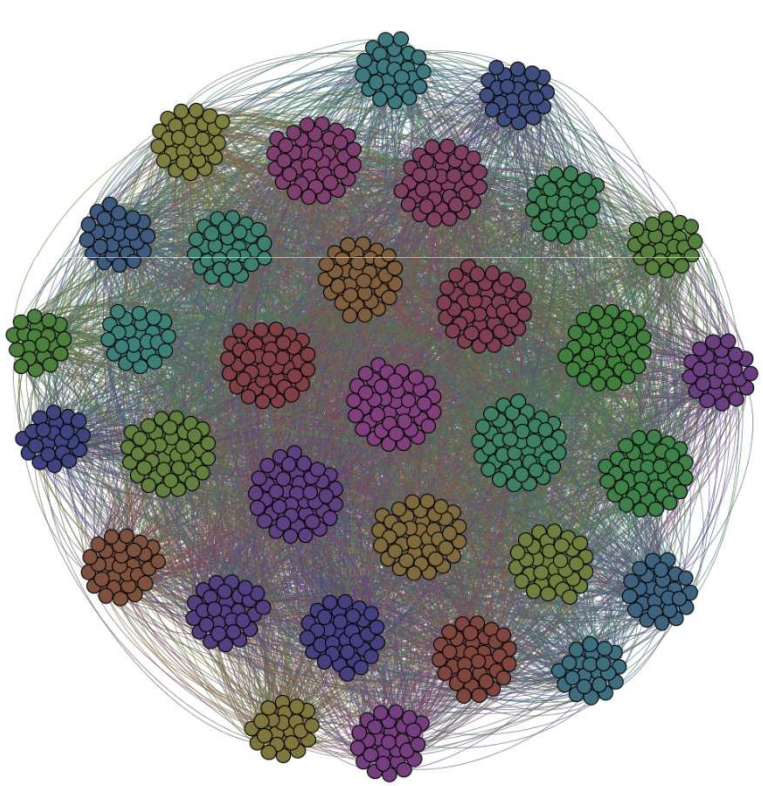
	Louvain	Label Prop.	SCCD
DBLP-top5K	0.53 (0.24)	0.49 (0.25)	0.51 (0.26)
Amazon-top5K	0.76 (0.37)	0.85 (0.46)	0.86 (0.46)
YouTube-top5K	0.23 (0.09)	0.08 (0.02)	0.19 (0.05)
Orkut-top5K	0.04 (0.03)	0.03 (0.02)	0.24 (0.06)
LJ-top5K	0.52 (0.28)	0.57 (0.32)	0.60 (0.32)
DBPedia-top5K	0.004 (0.0008)	0.003 (0.003)	0.13 (0.06)

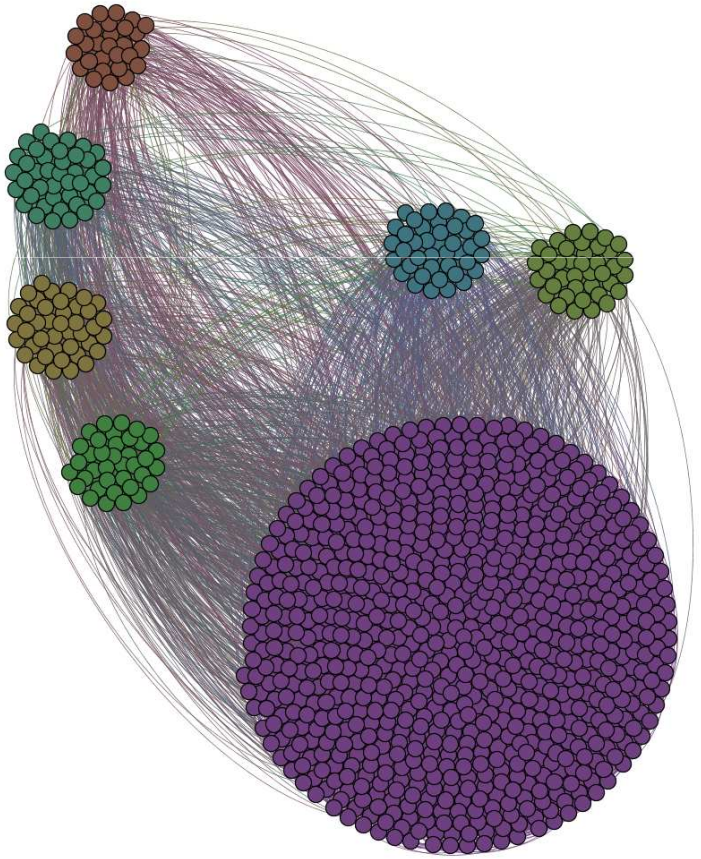
Detekcia komunít v rozsiahlych sieťach reálneho sveta

- Časť vrcholov prislúchajúcich 50-tim najväščím komunitám

	Lou fract. of N	LP fract. of N	SCCD fract. of N	Ground-truth fract. of N
DBLP	0.11	0.07	0.05	0.32
Amazon	0.03	0.03	0.04	0.83
YouTube	0.50	0.66	0.17	0.01
LJ	0.67	0.71	0.07	0.09
Orkut	0.97	0.99	0.93	0.09
DBPedia	0.98	0.96	0.89	0.001

Detekcia komunít v rozsiahlych sieťach reálneho sveta





Detekcia komunít v rozsiahlych sieťach reálneho sveta

- **Hypotéza:** tendencia skúmaných algoritmov založených na propagácii značiek na sieťach DBPedia a Orkut zhrnúť väčšinu vrcholov do malého počtu veľkých komunít je spôsobená veľkým podielom hrán spájajúcich väčšinu vrcholov s vrcholmi vysokého stupňa v hustom jadre siete.

- **Koeficient asortativity siete:** číselne vyjadruje tendenciu uzlov byť spojených s uzlami podobného stupňa (Pearsonov korelačný koeficient stupňov vrcholov spojených hranou)

Detekcia komunít v rozsiahlych sieťach reálneho sveta

- Asortativita skúmaných sietí a asortativita ich štruktúry explicitných komunít

	net AC	Top5k Comm. AC	All Comm. AC
DBLP	0.267	0.436	0.446
Amazon	-0.059	-0.077	-0.026
YouTube	-0.037	0.067	0.068
LJ	0.045	0.464	0.365
Orkut	0.016	0.233	0.326
DBPedia	-0.018	0.958	0.973

- Zistenie: Asortativita sietí a štruktúry komunít môže byť veľmi rozdielna.

Detekcia komunít v rozsiahlych sieťach reálneho sveta

- Skúmali sme možnosť modifikovať sieť pridaním váh na hrany
- Cieľom bolo znížiť váhu liniek spájajúcich disasortatívne uzly

Weighting function 1:

$$f_1(z,y) = floor(log_{10}(max(z,y)/min(z,y))$$

$$w1(e_{i,j}) = 10^{-f_1(d(i),d(j))}$$

Weighting function 2:

$$f_2(z,y) = floor(log_{10}(|z-y|)$$

$$w2(e_{i,j}) = 10^{-f_2(d(i),d(j))}$$

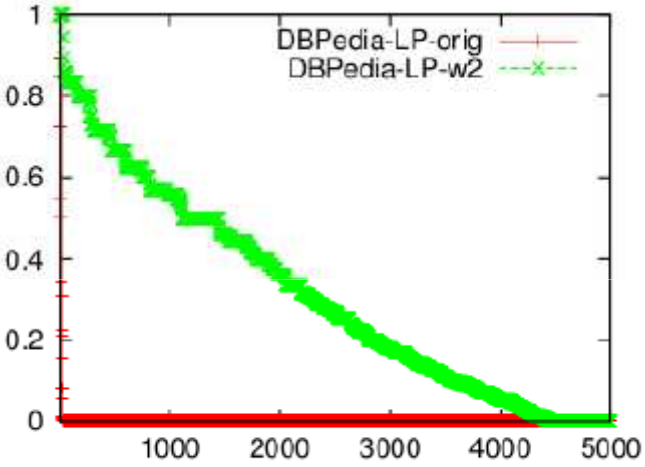
Detekcia komunít v rozsiahlych sieťach reálneho sveta

Network	Alg.	Orig	w1	w2
DBLP	Lou	0.54 (0.24)	0.54 (0.26)	**0.56** (0.26)
	LP	0.49 (0.25)	0.49 (0.24)	0.53 (0.25)
	SCCD	0.52 (0.26)	0.51 (0.25)	0.54 (0.25)
Amazon	Lou	0.76 (0.37)	0.84 (0.43)	0.83 (0.42)
	LP	0.85 (0.46)	0.85 (0.46)	0.83 (0.44)
	SCCD	**0.86** (0.46)	0.86 (0.46)	0.83 (0.44)
YouTube	Lou	0.23 (0.10)	0.23 (0.11)	**0.31** (0.12)
	LP	0.08 (0.02)	0.14 (0.05)	0.24 (0.09)
	SCCD	0.19 (0.06)	0.21 (0.08)	0.26 (0.10)
LJ	Lou	0.52 (0.28)	0.49(0.27)	0.52 (0.28)
	LP	0.57 (0.32)	0.59 (0.32)	0.62 (0.32)
	SCCD	0.6 (0.32)	0.61 (0.32)	**0.62** (0.33)
Orkut	Lou	0.04 (0.03)	0.04 (0.04)	0.17 (0.05)
	LP	0.03 (0.02)	0.03 (0.03)	0.11 (0.04)
	SCCD	0.24 (0.06)	0.20 (0.06)	**0.25** (0.06)
DBPedia	Lou	0.004 (0.0008)	0.016 (0.03)	0.053 (0.15)
	LP	0.003 (0.003)	0.054 (0.14)	0.31(0.17)
	SCCD	0.13 (0.06)	0.26 (0.15)	**0.34** (0.18)

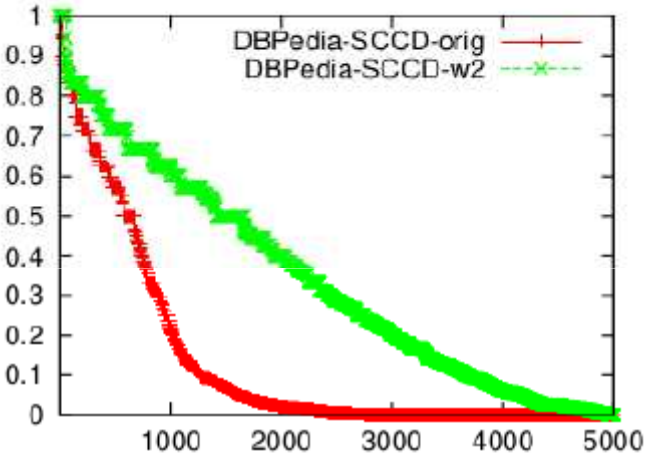
Detekcia komunít v rozsiahlych sieťach reálneho sveta

Net	Alg.	$w2$	rnd	MW	$MW \times w2$
DBLP	Lou	0.56	0.47(0.23)	0.54 (0.26)	0.56(0.26)
	LP	0.53	0.48(0.23)	0.56 (0.27)	0.56(0.26)
	SCCD	0.54	0.52(0.24)	0.57 (0.27)	**0.57**(0.27)
Amazon	Lou	0.83	0.83(0.43)	0.85 (0.43)	0.83(0.42)
	LP	0.83	0.82(0.43)	0.85 (0.44)	0.84(0.43)
	SCCD	0.83	0.82(0.42)	**0.85** (0.44)	0.83(0.42)
YouTube	Lou	0.31	0.15(0.05)	0.20 (0.09)	0.30(0.12)
	LP	0.24	0.11(0.04)	0.17 (0.07)	0.29(0.11)
	SCCD	0.26	0.21(0.06)	0.32 (0.13)	**0.32**(0.12)
LJ	Lou	0.52	0.50(0.27)	0.53 (0.28)	0.53(0.29)
	LP	0.62	0.56(0.31)	0.61 (0.33)	0.63(0.33)
	SCCD	0.62	0.62(0.33)	0.63 (0.34)	**0.63**(0.33)
Orkut	Lou	0.17	0.06(0.03)	0.08(0.03)	0.13(0.04)
	LP	0.11	0.03(0.02)	0.06(0.03)	0.18(0.04)
	SCCD	**0.25**	0.22(0.05)	0.16 (0.04)	0.19(0.05)
DBPedia	Lou	0.05	0.03(0.01)	0.05(0.03)	0.29(0.15)
	LP	0.31	0.001 (0.001)	0.33 (0.16)	0.40 (0.20)
	SCCD	0.34	0.15(0.06)	0.36 (0.18)	**0.41**(0.20)

Detekcia komunít v rozsiahlych sieťach reálneho sveta



Label propagation on DBPedia



SCCD on DBPedia

Výskumné smery

- Rozširovanie základného výskumu v algoritmoch
- Aplikácie detekcie komunít
- Škálovateľnosť

Aplikácie detekcie komunít

Aplikácie

- Analýza sociálnych sietí
  - **Rozdelenie do skupín**
- WWW – analýza liniek
  - **skupiny tématicky súvisiacich stránok**
- Biologické siete
  - **Interakčné siete proteínov – komunity zoskupujú proteíny rovnakej funkcie**
  - **Funkčné moduly v metabolických**
- Business inteligence
  - **Teleco – Churn prediction**
  - **Finančný sektor – fraud detection**

Aplikácie – Churn prediction v telekomunikáciách

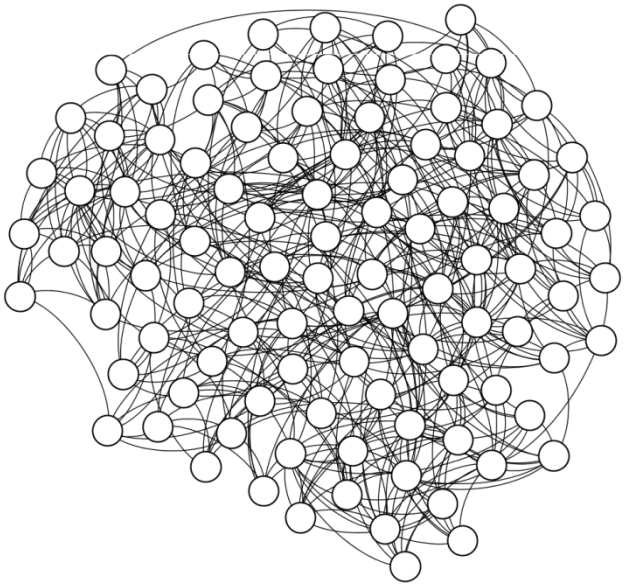
- Churn prediction
  - **Cieľ je predikovať možný odchod zákazníkov od mobilného operátora**
  - **Udržanie zákazníka lacnejšie ako získanie nového**

- Prediktívna analýza
  - **Dáta o zákazníkoch – vektory rysov**
  - **(f1, f2, ..., fn | churn=?)**

Aplikácie – Churn prediction v telekomunikáciách

- Churn prediction
  - **Cieľ je predikovať možný odchod zákazníkov od mobilného operátora**
  - **Udržanie zákazníka lacnejšie ako získanie nového**

- Prediktívna analýza
  - **Dáta o zákazníkoch – vektory rysov**
  - **(f1, f2, ..., fn | churn=?)**
  - **Rozšírenie vektorov o grafové rysy**
    - Príslušnosť do komunity
    - Centralita uzla (napr. page rank)
    - ....
  - **(f1, f2, ..., fn_, gf1, gf2, ..., gf_m | churn=?)**



Škálovateľnosť

- Motivácia
  - **Sociálna sieť – relácie medzi užívateľmi**
    - Facebook - 750 miliónov užívateľov
    - YouTube - 490 miliónov pravidelných užívateľov
    - Twitter - 550 miliónov užívateľov
    - Wikipedia - 91,000 kontribútorov
  - **Obsah generovaný užívateľmi**
    - textové správy, fotky, videá, reakcie (+1 / Likes)
    - Facebook - 30 miliárd zdieľaných položiek / mesiac
    - Twitter - 190 miliónov mikropostov / deň
    - Wikipedia - 17 miliónov hostovaných článkov
  - **Interakcia užívateľov s obsahom**
    - YouTube - 92 miliárd zobrazení stránok / mesiac
    - Twitter - 1.6 miliárd dopytov za deň
    - Facebook – priemerný čas strávený na stránke za mesiac: 15H 33M

Škálovateľnosť

- Výzvy:
  - **Časová náročnosť algoritmov**
  - **Pamäťové ohraničenie**

- Pamäťové ohraničenie
  - **Distribuované rámce pre výpočty nad grafmy**
  - **Koncept „Pregel“ od googlu**

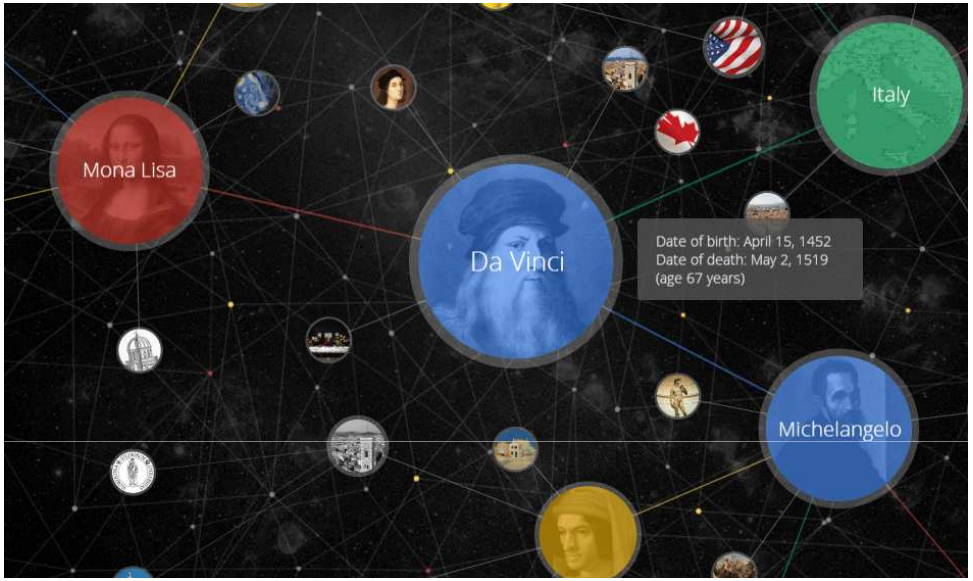
    [Grzegorz Malewicz, Matthew H. Austern, Aart J.C Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. 2010. Pregel: a system for large-scale graph processing. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10). ACM, New York, NY, USA, 135-146. DOI=10.1145/1807167.1807184]
  - **Open source implementácie**
    - Giraph, Apache Hama GraphLab

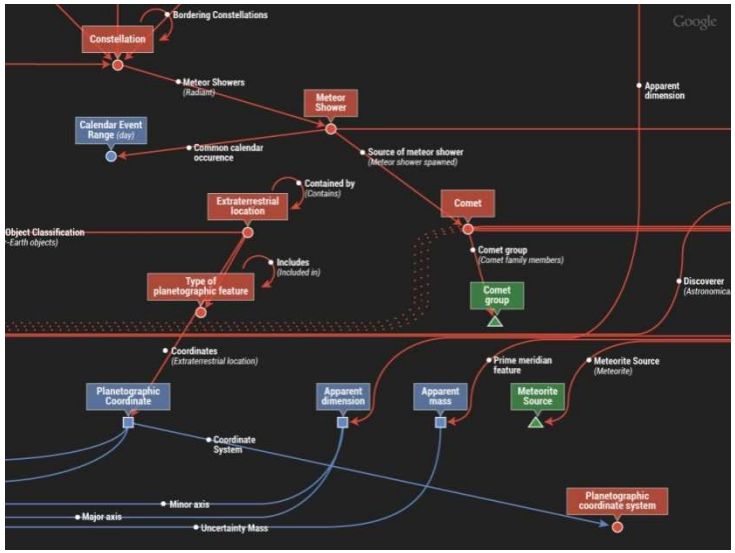
Applications of Graphs and Networks

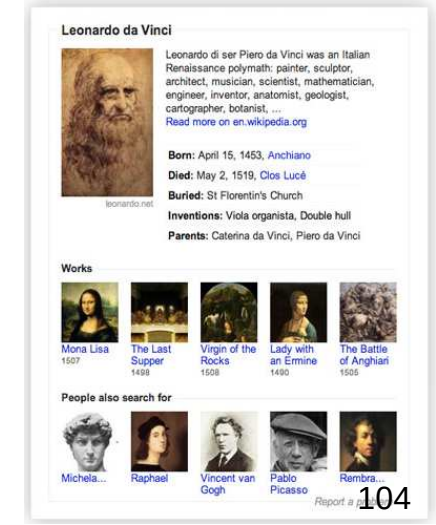
Google Knowledge Graph

[ulanoff]

- Wikipedia
- Freebase
- Confirmed human knowledge



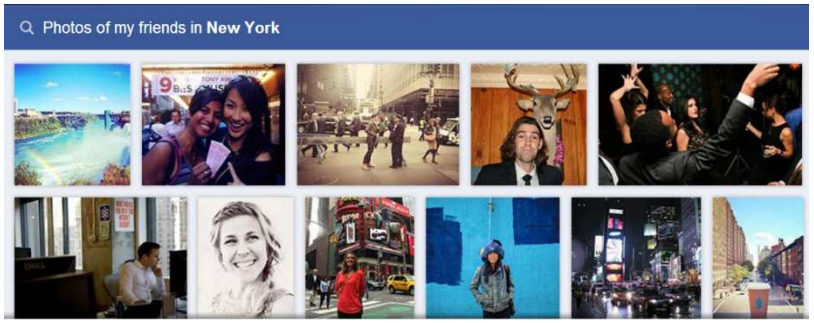


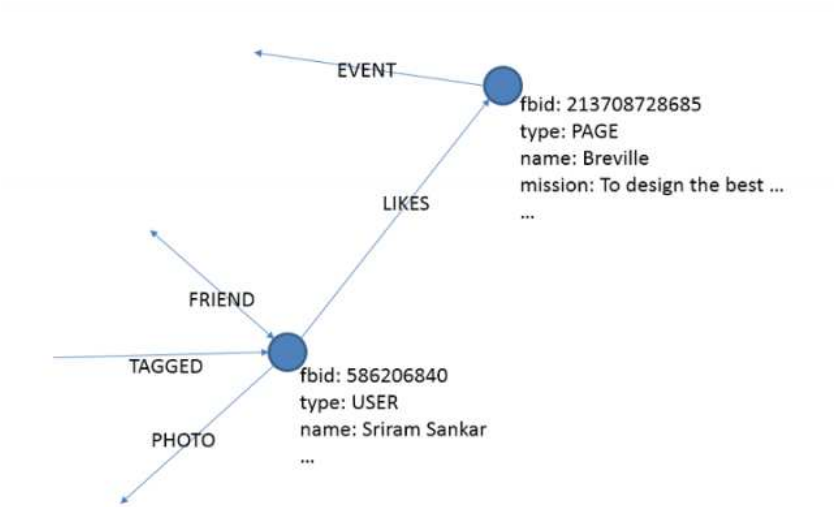


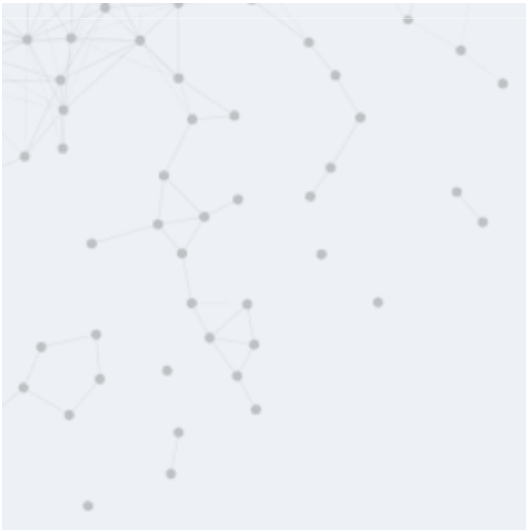
Facebook Graph Search

[facebook13]

- Užívateľmi generovaný obsah
- Prepojenia na web

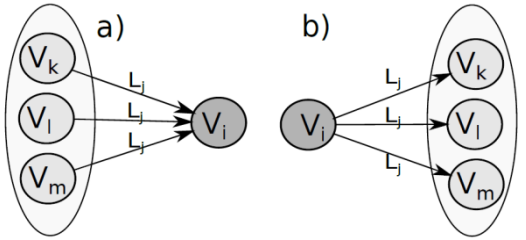




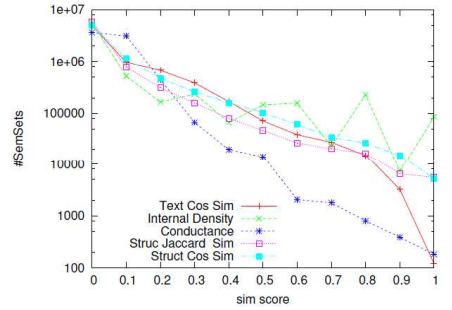


SemSets: Sematnic Search

- Answering list type questions: *astronauts who walked on the Moon*
- Wikipedia as text and networks/graph
- Text: IR methods, Lucene based
- Graph/network: sprading activation and SemSets
- Winning solution on Semantic Search Challenge 2011



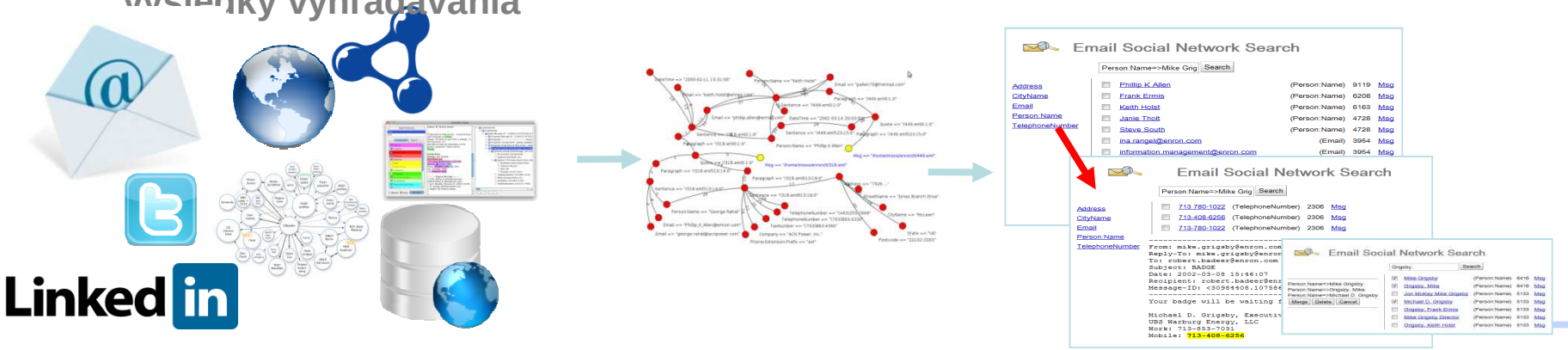
[SemSets]



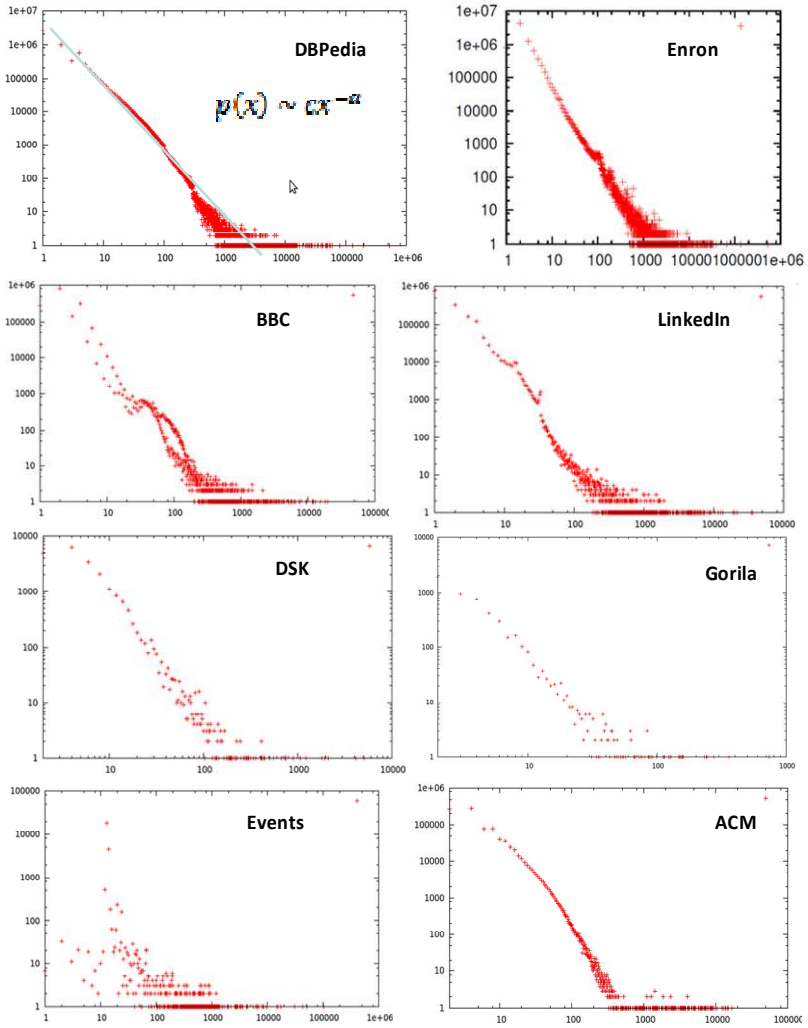
1. Eugene_Cernan
2. Alan_Bean
3. David_Scott
4. John_Young_(astronaut)
5. Neil_Armstrong
6. Pete_Conrad
7. Harrison_Schmitt
8. Alan_Shepard
9. Charles_Duke
10. Buzz_Aldrin
11. James_Irwin
12. Edgar_Mitchell

Objavovanie vzťahov vo veľkých grafových dátach

- Motivácia
  - **Grafy a siete sú všadeprítomné : sociálne site, web, LinkedData, transakcie, komunikácia (email, telefóny).**
  - **Text tiež môže byť prevedený na graf.**
  - **Prepojenie grafových dát a vyhľadávania relácii v nich je dôležite**
- Prístup
  - **Tvorba sémantických stromov a grafov z textu, webu, komunikácie, databáz a LinkedData**
  - **Užívateľská interakcia s týmito dátami aby sa dali lepšie integrovať zdroje a vyčistiť upraviť dáta**
  - **Užívatelia to budú robiť ak to bude mať zmysel, teda okamžitý vplyv na lepšie výsledky vyhľadávania**



Vlastností vybraných grafov/sietí



Názov siete	Počet vrcholov	Počet hrán	Priem. klást. koef.	Koef. assort.	Priem. najkr. cesta
Enron Full	8 269 278	20 383 709	0,29	-0,02	6,58
Enron5	160 387	630 330	0,30	-0,04	6,64
LinkedIn	1 564 698	6 094 634	0,36	0,13	6,48
BBC	1 725 900	6 839 358	0,34	-0,05	7,55
DSK	21 518	98 952	0,31	0,39	5,79
DSK3	2 857	8 754	0,36	-0,14	5,46
Gorila	5 959	23 724	0,31	0,03	6,25
Events	25 478	539 328	0,38	-0,25	2,47
ACM	941 322	2 198 001	0,34	-0,06	7,30

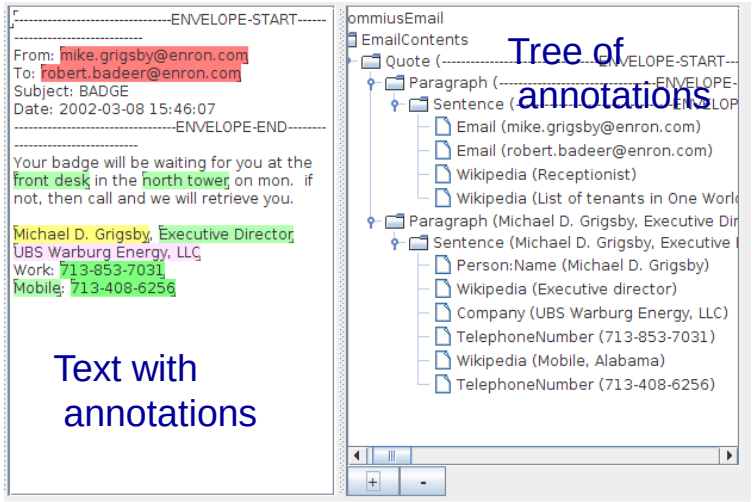
Datasety:

- DBPedia
- Web
  - BBC, LinkedIn, DSK
- Gorila – document
- Events – agent simulation event graph
- ACM – publications, LinkedData

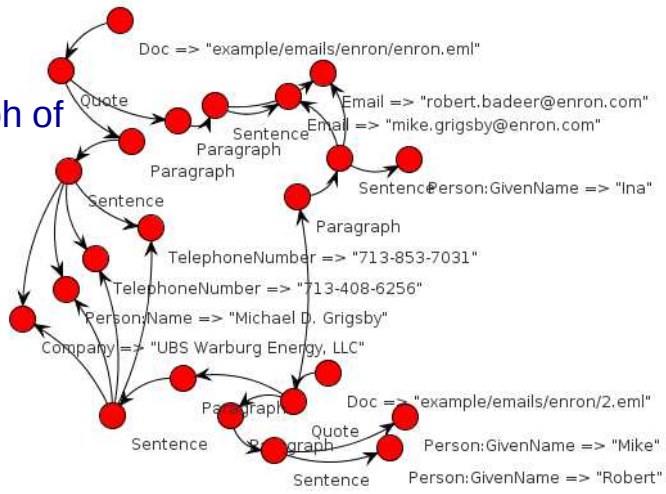
Ontea: Information Extraction Tool

http://ontea.sf.net

- Regex patterns
- Gazetteers
- Resuls
  - Key-value pairs
  - Structured into trees
  - graphs
- Transformers, Configuration
- Automatic loading of extractors



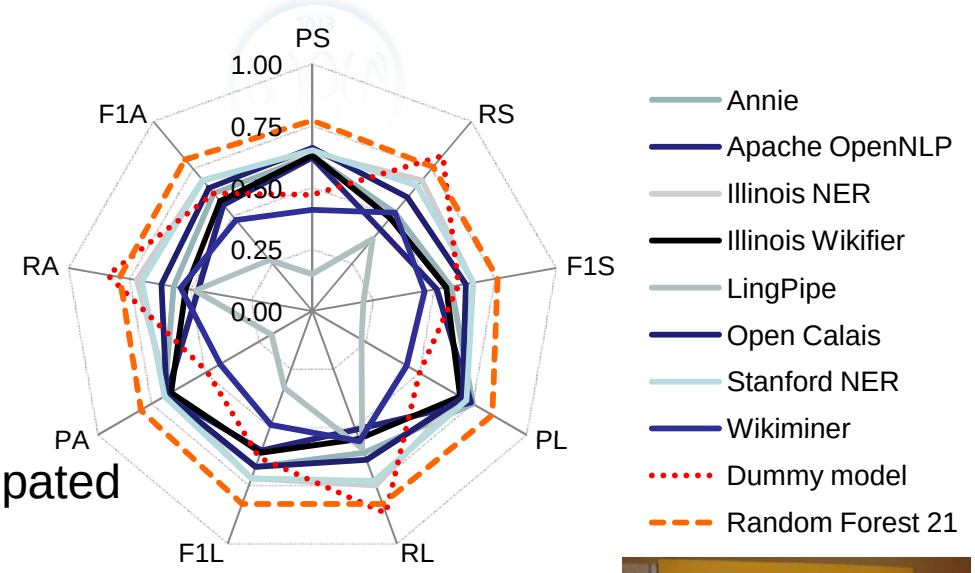
- Visual Annotation Tool
- Integration with external tools
  - GATE, Stemers, Hadoop …
- Multilingual tests
  - **English, Slovak,** Spanish, Italian

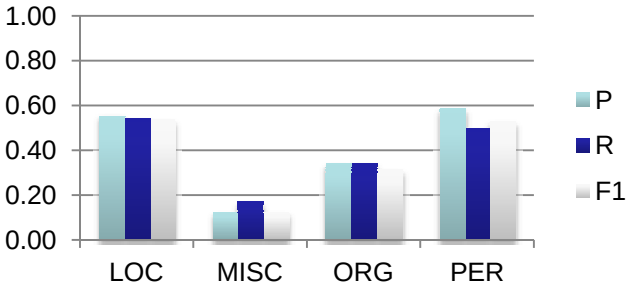


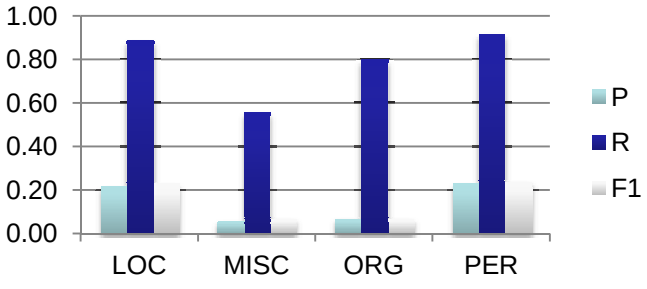
Named Entity Recognition (NER)

- Combination of Existing NER
  - ANNIE (GATE), Apache OpenNLP,
  - Illinois NER, Illinois Wikifier,
  - LingPipe, Open Calais
  - Stanford NER ,WikiMiner,
  - Miscinator
- Machine Learning
  - Decision Trees models
- Received second place at MSM 2013, missing first place by 1%, where participated 17 teams world wide
  http://ikt.ui.sav.sk/index.php?n=Main.IEChallenge2013

Micro Summary (test set)





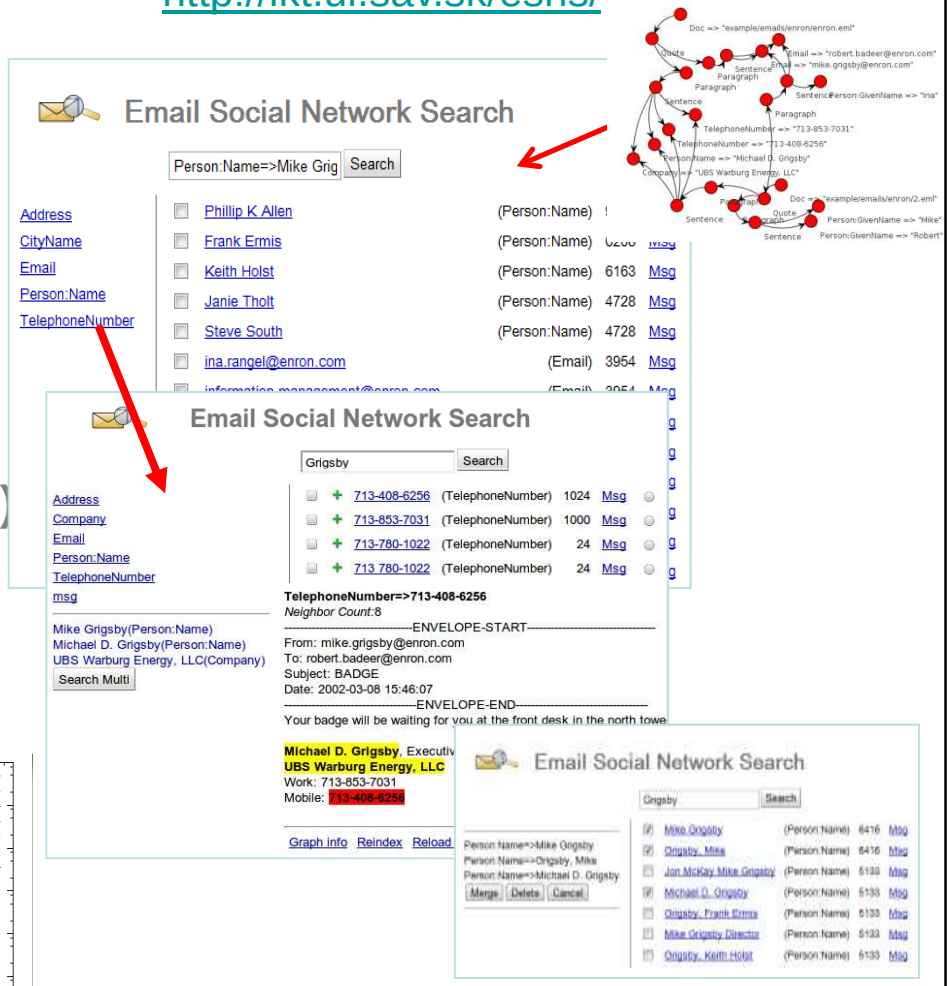


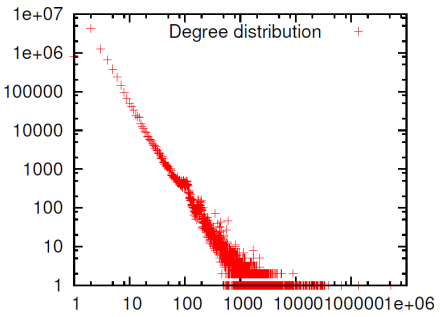
gSemSearch: Graph based Semantic Search

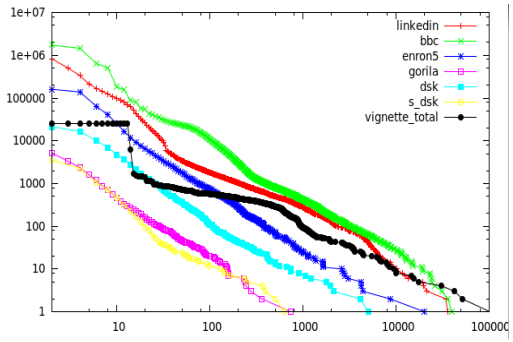
[gSemSearch]

http://ikt.ui.sav.sk/esns/

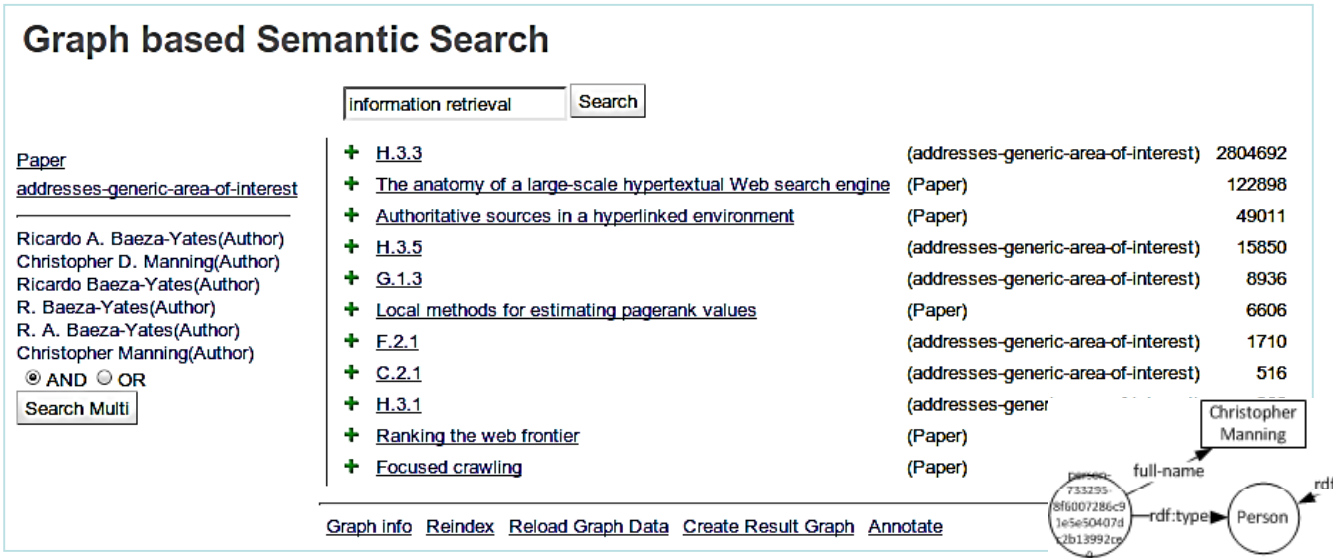
- Entity relation search in semantic networks/graphs
- Search, Navigation, Data Interaction
- Aiming at data integration of
  - **Structured data (Relational data, LinkedData)**
  - **Unstructured Data (text, documents, communication)**
- Applications:
  - **Email, Web, Text documents, LinkedData**

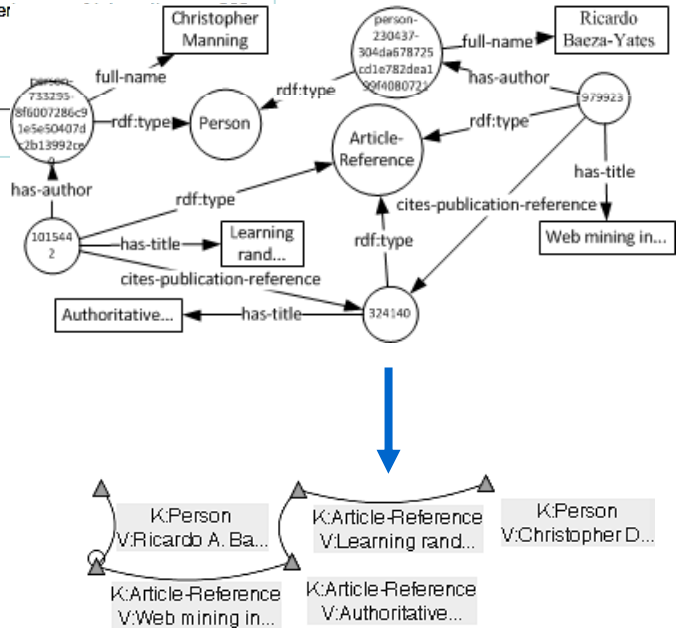






Navigácia v zjednodušenom LinkedData grafe

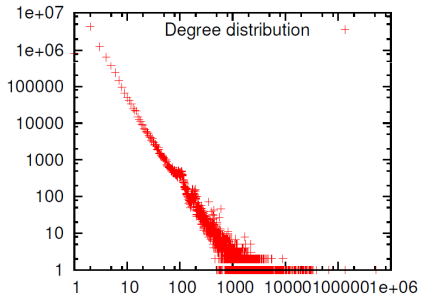


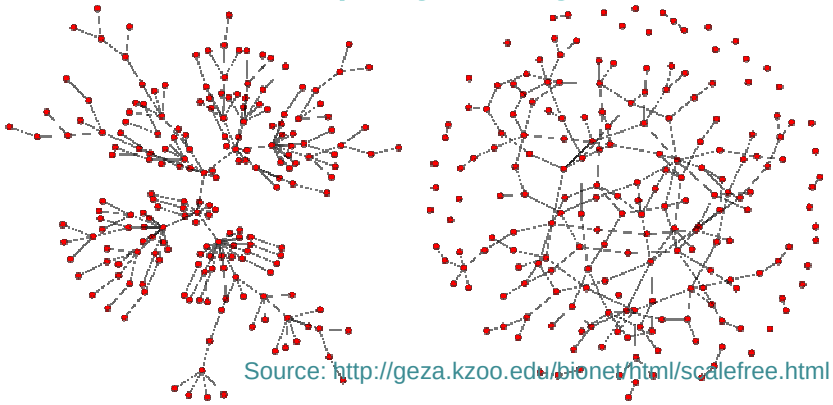


- Konverzia ACM LinkedData na jednoduchý graf pre gSemSearch
  - **Experiment na hľadanie relácií a navigáciu**
  - **Pri konverzii na jednoduchší graf zanedbanie typov vzťahov: niekedy problém**

SGDB: Simple Graph Database

- Storage for graphs
- Optimized for graph traversing and spread of activation
- Faster then Neo4j for graph traversing operations
- Supports Blueprints API
- https://simplegdb.svn.sourceforge.net/svnroot/simplegdb/Sgdb3





Source: http://geza.kzoo.edu/~bionet/html/scalefree.html

- Graph Database Benchmarks
  - **Graph Traversal Benchmark for Graph Databases**
  - **http://ups.savba.sk/~marek/gbench.html**
  - **Blueprints API - possibility to test compliant Graph databases**

IBM Watson

[Perrone11]

IBM Research

IBM

Informed Decision Making: Search vs. Expert Q&A

Decision Maker

- Has Question
- Distills to 2-3 Keywords
- Reads Documents, Finds Answers
- Finds & Analyzes Evidence

Search Engine

- Finds Documents containing Keywords
- Delivers Documents based on Popularity

Expert

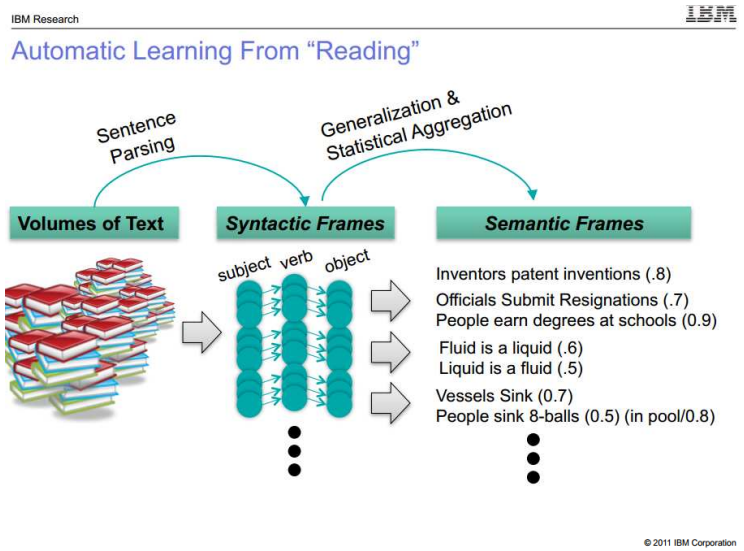
- Understands Question
- Produces Possible Answers & Evidence
- Analyzes Evidence, Computes Confidence
- Delivers Response, Evidence & Confidence

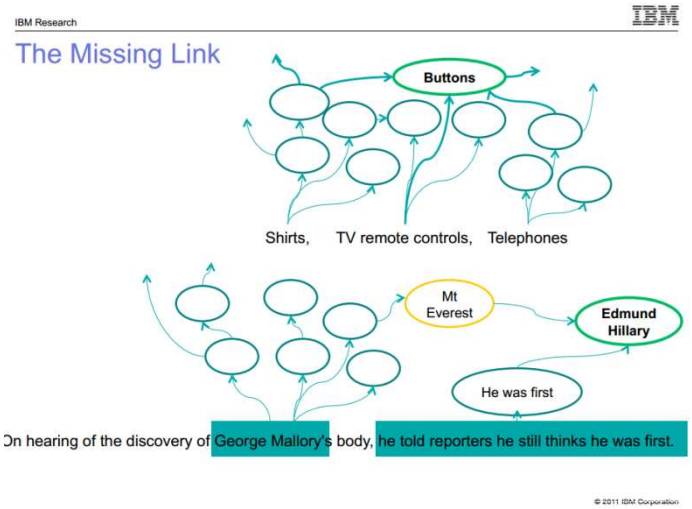
Decision Maker

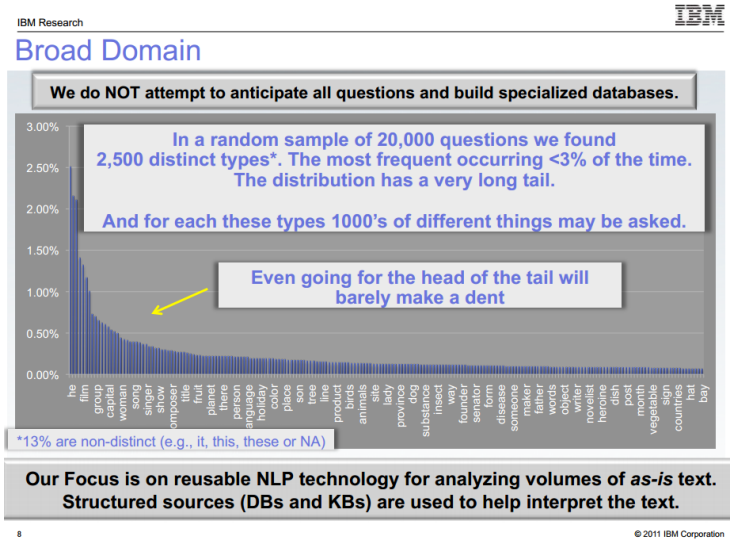
- Asks NL Question
- Considers Answer & Evidence

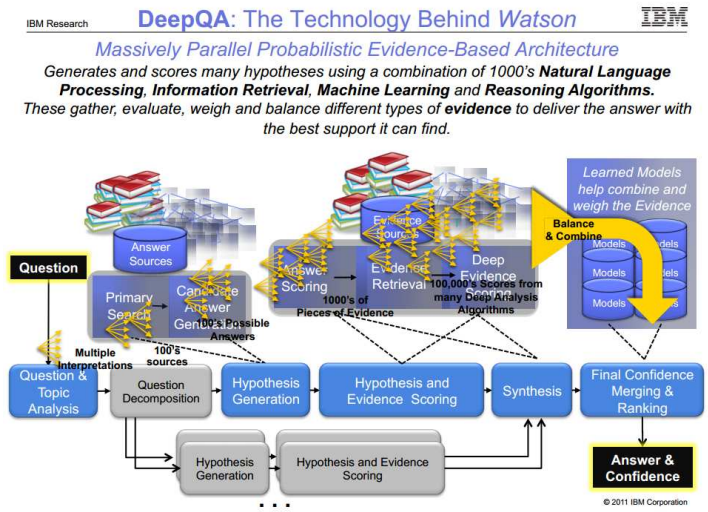
IBM Watson

[Perrone11]



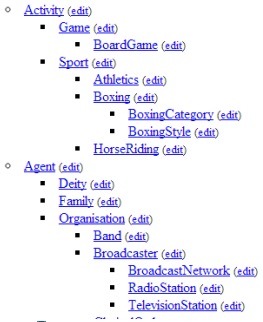






Data Sources

- Wikipedia
  - 6 million articles
  - 40 GB text
- DBPedia
  - Triples
  - Good categories for articles
- Freebase
  - 170 GB triples
  - 40 million topics
  - 1.2 billion triples



Smartphone /m/0169zh

Video Game Platform, Computing Platform, Literature Subject, Product category, Competitive Space

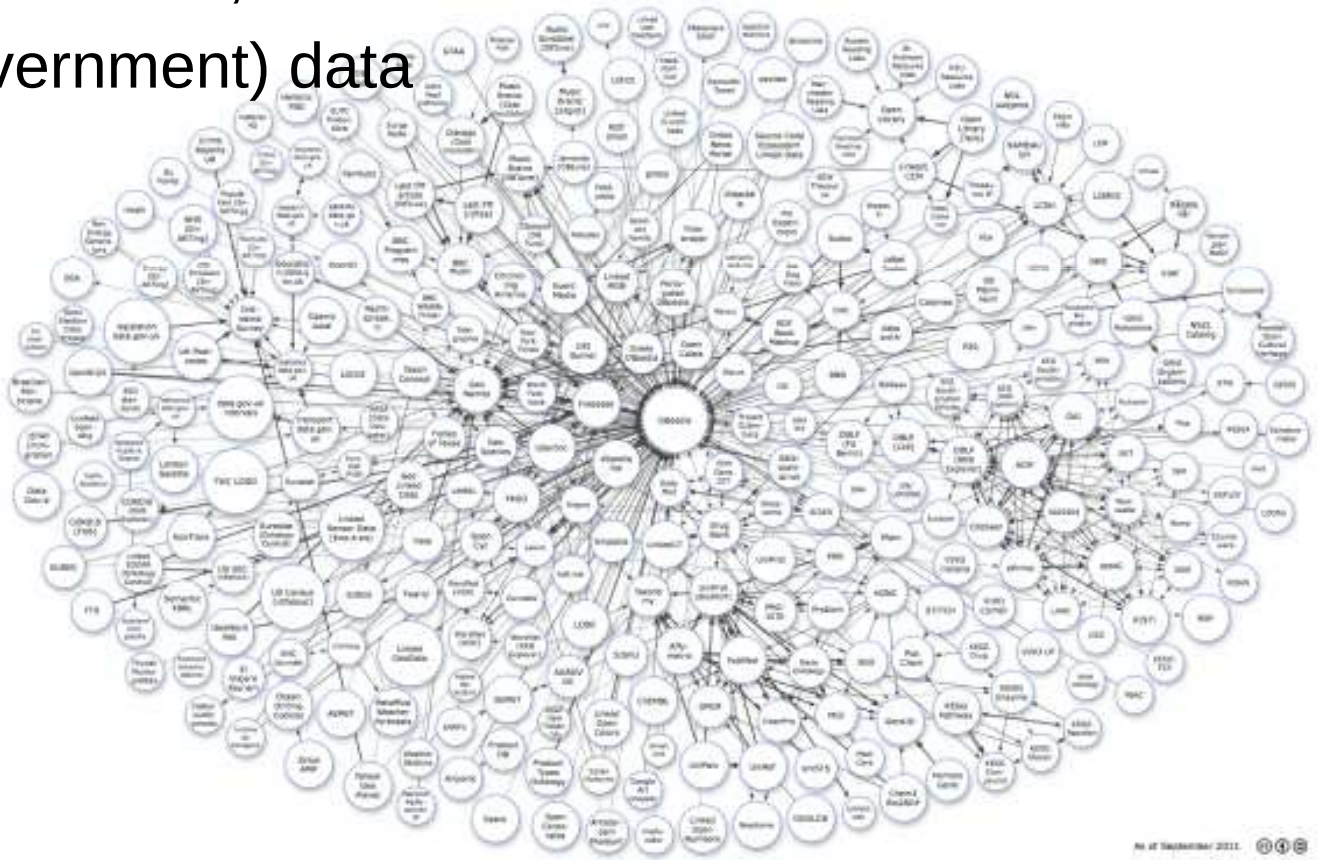
alias: smart phone, smartphone

A smartphone, or smart phone, is a mobile phone built on a mobile operating system, with more advanced computing capability and connectivity than a feature phone. The first smartphones combined the functions of a personal digital assistant with a... [Wikipedia]

```
ns:m.012rkqx    ns:type.object.type ns:common.topic.
ns:m.012rkqx    ns:type.object.name "High Fidelity"@en.
ns:m.012rkqx    ns:type.object.type ns:music.single.
ns:m.012rkqx    ns:type.object.key  ns:authority.musicbrainz.name.TRACK3987054.
ns:m.012rkqx    ns:type.object.type ns:music.recording.
ns:m.012rkqx    key:authority.musicbrainz   "258c45bd-4437-4580-8988-b3f3be975f9c".
ns:m.012rkqx    key:authority.musicbrainz.name  "TRACK3987054".
ns:m.012rkqx    rdfs:label  "High Fidelity"@en.
ns:m.012rkqx    rdfs:type   ns:common.topic.
ns:m.012rkqx    rdfs:type   ns:music.single.
ns:m.012rkqx    rdfs:type   ns:music.recording.
```

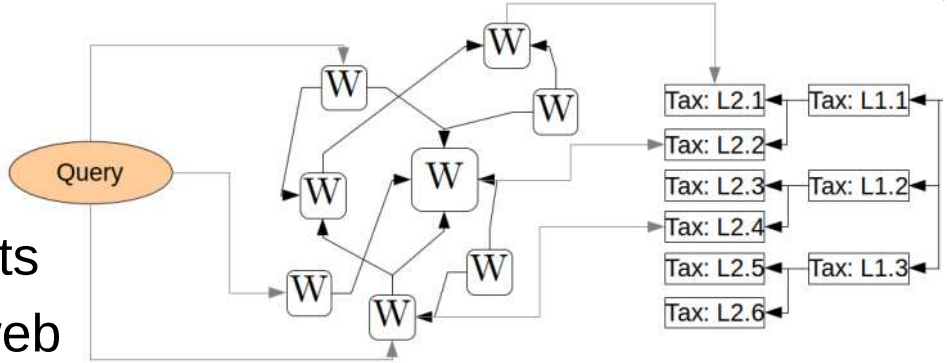
Linked Data cloud

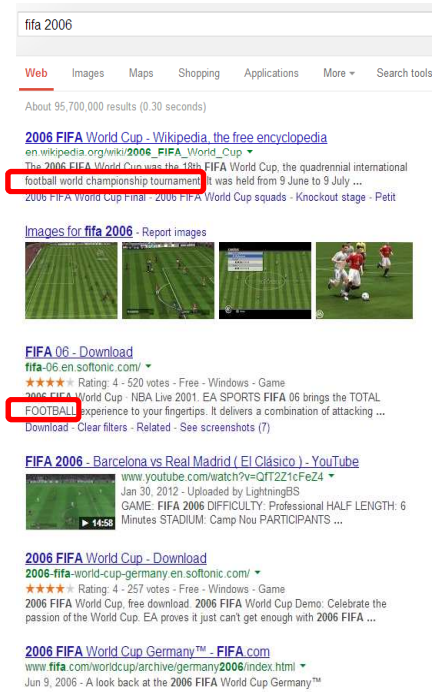
- Triples (graph)
- DBPedia, Geo, people, publications, medicine, …
- EU public (government) data



Query Categorization (QC)

- Usual approach for QC:
  - Get results for query
  - Categorize returned documents
- Best algorithms work with entire web (search API)





Query	Categories
apple	Computers \ Hardware Living \ Food & Cooking
FIFA 2006	Sports \ Soccer Sports \ Schedules & Tickets Entertainment \ Games & Toys
cheesecake recipes	Living \ Food & Cooking Information \ Arts & Humanities
friendships poem	Information \ Arts & Humanities Living \ Dating & Relationships

Representing categories by Wiki pages

- Each category represented by one or multiple wiki pages
- Example: *IAB19-6 Cell Phones* category
  - **Mobile phone; Smartphone; Camera phone.**

IAB1 Arts & Entertainment		The arts	Entertainment			
IAB1-1 Books & Literature		Book	Literature	Novel		
IAB1-2 Celebrity Fan/Gossip		Celebrity				
IAB1-3 Fine Art	Fine art					
IAB1-4 Humor	Humour	Fun				
IAB1-5 Movies	Film	Filmmaking	Film industry	Film genre	Movie theater	
IAB1-6 Music	Music	Pop music	Classical music	Rock music	Music genre	List of popular music genres
IAB1-7 Television	Television program	Television	Serial (radio and television)			
01-03 Performing art & Theatre		Performing arts	Theatre			
IAB2 Automotive	Automotive industry	Automobile	Car classification			
IAB2-1 Auto Parts	List of auto parts					
IAB2-2 Auto Repair	Automotive Service Excellence					
IAB2-3 Buying/Selling Cars		Used car	Car dealerships in North America			
IAB2-4 Car Culture	Effects of the automobile on societies					
IAB2-5 Certified Pre-Owned		Certified Pre-Owned				
IAB2-6 Convertible	Convertible					
IAB2-7 Coupe	Coupé	Compact car				
IAB2-8 Crossover	Crossover (automobile)					
IAB2-9 Diesel	Turbo-diesel					
IAB2-10 Electric Vehicle		Electric car				
IAB2-11 Hatchback	Hatchback					
IAB2-12 Hybrid	Hybrid vehicle					
IAB2-13 Luxury	Luxury vehicle					
IAB2-14 MiniVan	Minivan					
IAB2-15 Morорcycles	Motorcycle	List of motorcycle manufacturers				
IAB2-16 Off-Road Vehicles		Off-road vehicle				
IAB2-17 Performance Vehicles		Performance car				
IAB2-18 Pickup	Pickup truck					
IAB2-19 Road-Side Assistance		Roadside assistance	AAA (American Automobile Association)			
IAB2-20 Sedan	Sedan (automobile)					
IAB2-21 Trucks & Accessories		Truck	Truck accessory			
IAB2-22 Vintage Cars	Vintage car	History of the automobile				
IAB2-23 Wagon	Station wagon					
02-03 Campers & RVs	Truck camper	Recreational vehicle				
02-06 Concept Cars	Concept car					
02-08 SUV	Sport utility vehicle					


Mobile phone

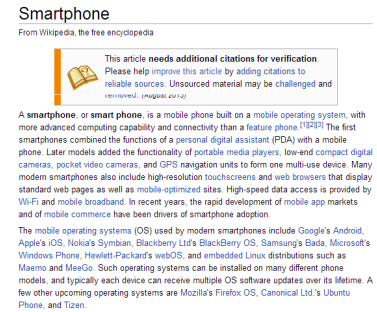
From Wikipedia, the free encyclopedia

"Cell Phone" redirects here. For the film, see Cell Phone (film). For the Handphone film, see Handphone (film).

A **mobile phone** (also known as a **cellular phone**, **cell phone**, and a **hand phone**) is a device that can make and receive telephone calls over a radio link while moving around a wide geographic area. It does so by connecting to a cellular network provided by a mobile phone operator, allowing access to the public telephone network. By contrast, a cordless telephone is used only within the short range of a single, private base station.

In addition to telephony, modern mobile phones also support a wide variety of other services such as text messaging, MMS, email, Internet access, short-range wireless communications (infrared, Bluetooth), business applications, gaming and photography. Mobile phones that offer these and more general computing capabilities are referred to as smartphones.

The Qualcomm QCP-2700, a mid-1990s candybar style phone, and an iPhone 5, a current production smartphone.


Smartphone

From Wikipedia, the free encyclopedia

This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and

A **smartphone**, or **smart phone**, is a mobile phone built on a mobile operating system, with more advanced computing capability and connectivity than a feature phone.[1][2][3] The first smartphones combined the functions of a personal digital assistant (PDA) with a mobile phone. Later models added the functionality of portable media players, low-end compact digital cameras, pocket video cameras, and GPS navigation units to form one multi-use device. Many modern smartphones also include high-resolution touchscreens and web browsers that display standard web pages as well as mobile-optimized sites. High-speed data access is provided by Wi-Fi and mobile broadband. In recent years, the rapid development of mobile app markets and of mobile commerce have been drivers of smartphone adoption.

The mobile operating systems (OS) used by modern smartphones include Google's Android, Apple's iOS, Nokia's Symbian, Blackberry Ltd's BlackBerry OS, Samsung's Bada, Microsoft's Windows Phone, Hewlett-Packard's webOS, and embedded Linux distributions such as Maemo and MeeGo. Such operating systems can be installed on many different phone models, and typically each device can receive multiple OS software updates over its lifetime. A few other upcoming operating systems are Mozilla's Firefox OS, Canonical Ltd.'s Ubuntu Phone, and Tizen.


Camera phone

From Wikipedia, the free encyclopedia

For the song by The Game, see Camera Phone (song).

See also: Mobile phone and Videophone

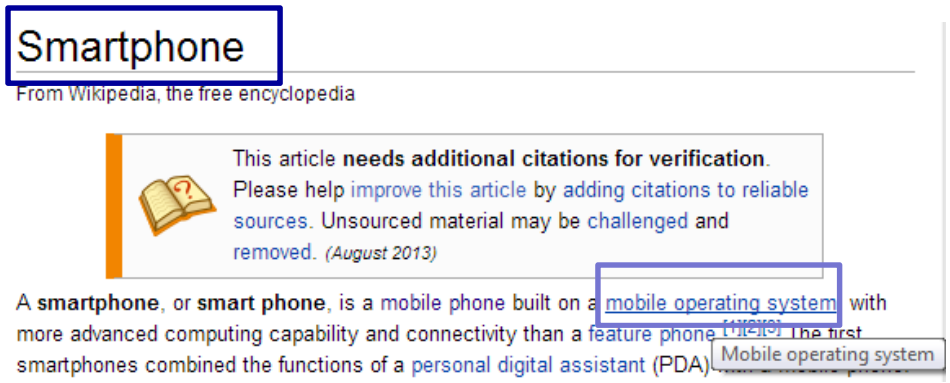
This article's factual accuracy may be compromised due to out-of-date information. Please update this article to reflect recent events or newly available information. (May 2013)

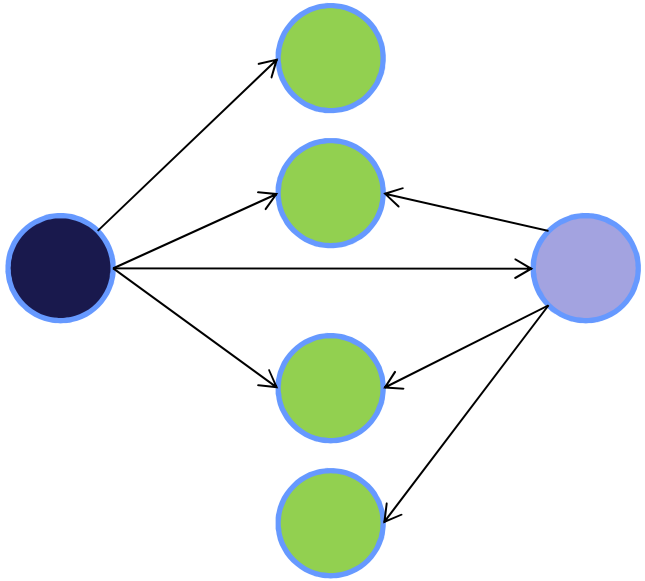
The camera phone solution allows instant sharing of pictures. As it's automatic and instant, the user does not have to use a cable or removable media to connect to a personal computer.

A **camera phone** is a mobile phone which is able to capture still photographs (and usually video). Since early in the 21st century the majority of mobile phones in use are camera phones.[1]

Kewords / Ngrams Generation From Wikipedia

- Seed concpets – manually assigned for each category
- Exploit Wiki link graph
- For all neigboring Wiki Concepts (to seed concepts)
  - **Filter out concepts of type: Person, Location, Org, Work**
  - **Compute cosine similarity of adjedcemcy lists**
  - **Extend concept names by alternative names (redirect pages from Wiki)**





Literatúra

- [Ulanoff] Lance Ulanoff: **Google Knowledge Graph Could Change Search Forever** http://mashable.com/2012/02/13/google-knowledge-graph-change-search/, 2012
- [facebook13] Sean Gallagher, **Knowing the score: How Facebook's Graph Search knows what you want,** http://arstechnica.com/information-technology/2013/03/knowing-the-score-how-facebooks-graph-search-knows-what-you-want/, 2013
- [Perrone11] Michael Perrone: **What is Watson – An Overview**, 2011, http://static.usenix.org/event/lisa11/tech/slides/perrone.pdf
- [WatsonJr] Tony Pearson: **IBM Watson - How to build your own "Watson Jr." in your basement**, 2012, https://www.ibm.com/developerworks/mydeveloperworks/blogs/InsideSystemStorage/entry/ibm_watson_how_to_build_your_own_watson_jr_in_your_basement7?lang=en
- [OpenNLP] OpenNLP: http://www.slideshare.net/gagan1667/opennlp-demo
- [TamingText] Ingersoll, G., Morton, T., & Farris, L. (2012). Taming Text: How to find, organize and manipulate it.
- [Zaragoza] Hugo Zaragoza. **Machine Learning and Information Retrieval**, ESSIR 2009 Lecture
- [Nigam] Kamal Nigam: **Generative Models for Text Classification and Information Extraction**, http://www.cs.cmu.edu/~knigam/15-505/ie-lecture.ppt

Literatúra

- [SemSets] CIGLAN, Marek - NoRVaG, Kjetil - HLUCHÝ, Ladislav. **The SenSets model for ad-hoc semantic list search**. In WWW´12 Proceedings of the 21st International Conference on World Wide Web. - New York : ACM, 2012, p. 131-140. ISBN 978-1-4503-1229-5. SCOPUS, http://www2012.wwwconference.org/proceedings/proceedings/p131.pdf
- [gSemSearch] LACLAVÍK, Michal - DLUGOLINSKÝ, Štefan - ŠELENG, Martin - CIGLAN, Marek - HLUCHÝ, Ladislav. **Emails as graph: relation discovery in email archive**. In WWW´12 Companion Proceedings of the 21st International Conference companion on World Wide Web. - New York : ACM, 2012, 841-846. ISBN 978-1-4503-1230-1. http://www2012.wwwconference.org/proceedings/companion/p841.pdf . SCOPUS
- [gBench] CIGLAN, Marek - AVERBUCH, Alex - HLUCHÝ, Ladislav. **Benchmarking traversal operations over graph databases**. In 2012 IEEE 28th International Conference on Data Engineering Workshops : proceedings. - Los Alamitos : IEEE Computer Society, 2012, p. 186-189. ISBN 978-1-4673-1640-8. SCOPUS
- [ontea_email] LACLAVÍK, Michal - DLUGOLINSKÝ, Štefan - ŠELENG, Martin - KVASSAY, Marcel - GATIAL, Emil - BALOGH, Zoltán - HLUCHÝ, Ladislav. **Email analysis and information extraction for enterprise benefit. In Computing and informatics**, 2011, vol. 30, no. 1, p. 57-87. (0.356 - IF2010). ISSN 0232-0274.
- [uiWeb] Dlugolinský, Štefan - Šeleng, Martin - Laclavík, Michal - Hluchý, Ladislav. **Distributed Web-scale Infrastructure for Crawling, Indexing and Search with Semantic Support**. In Computer Science Journal, 13 (4)

Ďakujeme za pozornosť